The Whys and Hows of Data Augmentation

DMQA Open Seminar

Hyungu Kahng

Department of Industrial and Management Engineering
Korea University

January 22, 2021

발표자: 강현구

- 학력
  - 2011.03 - 2015.02: 학사, 고려대학교 산업경영공학과
  - 2015.03 – 현재: 석박사통합과정, 고려대학교 산업경영공학과 (지도교수: 김성범)
- 연구분야
  - Self-supervised visual representation learning and its industrial applications
  - Deep reinforcement learning algorithms for real-time strategy games
  - Generative models for missing data imputation
  - Machine learning applications for medical data analysis

Outline

1) Why is data augmentation necessary?

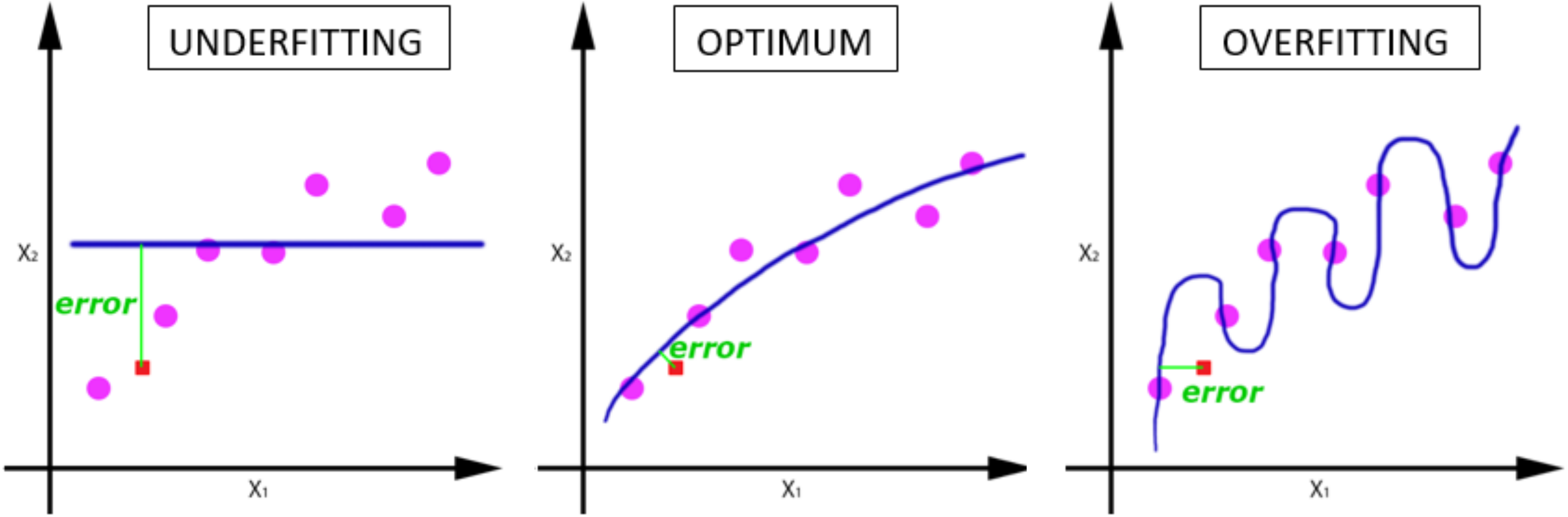
2) How is data augmentation done?

   - in computer vision
   - in natural language processing

Overfitting

A common problem in machine learning (regression example)

- The model performs well on training data (●) but generalizes poorly to unseen data (■).
- Ideally, a sweet spot always exists.

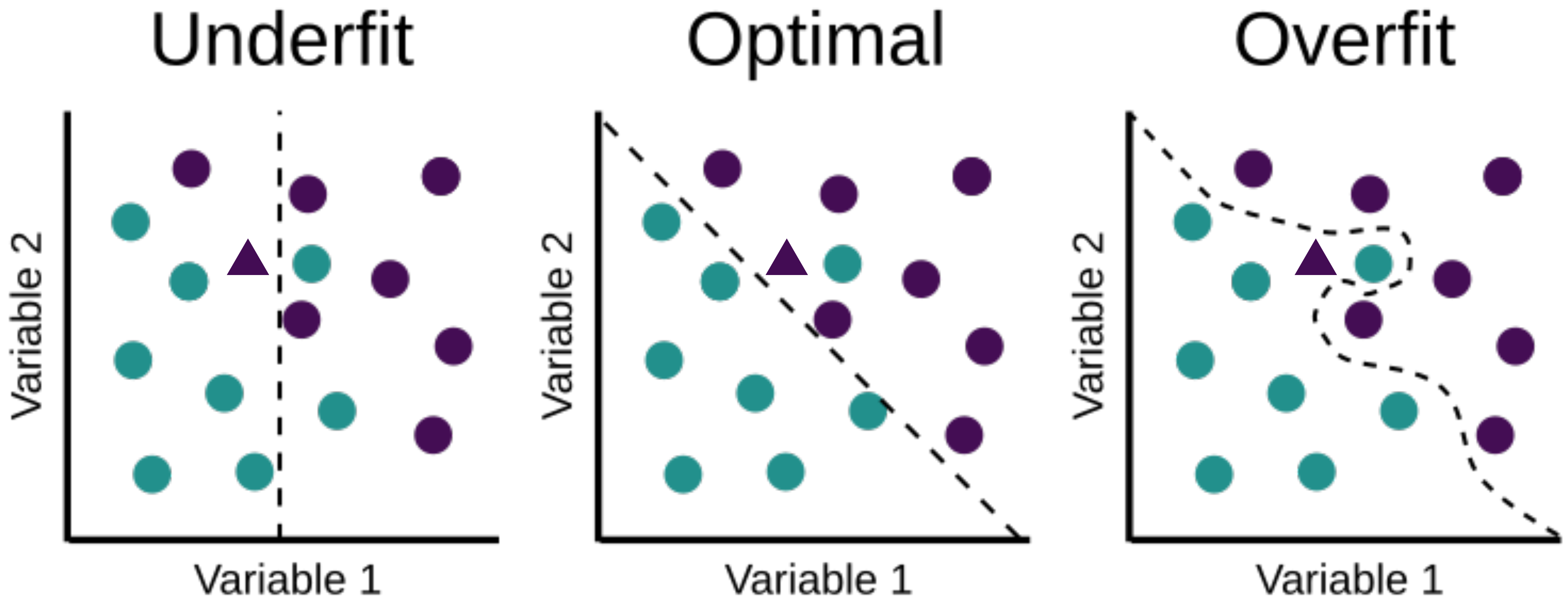


https://docs.paperspace.com/machine-learning/wiki/overfitting-vs-underfitting

Overfitting

A common problem in machine learning (classification example)

- The model performs well on training data (●●) but generalizes poorly to unseen data (▲).
- Ideally, a sweet spot always exists.

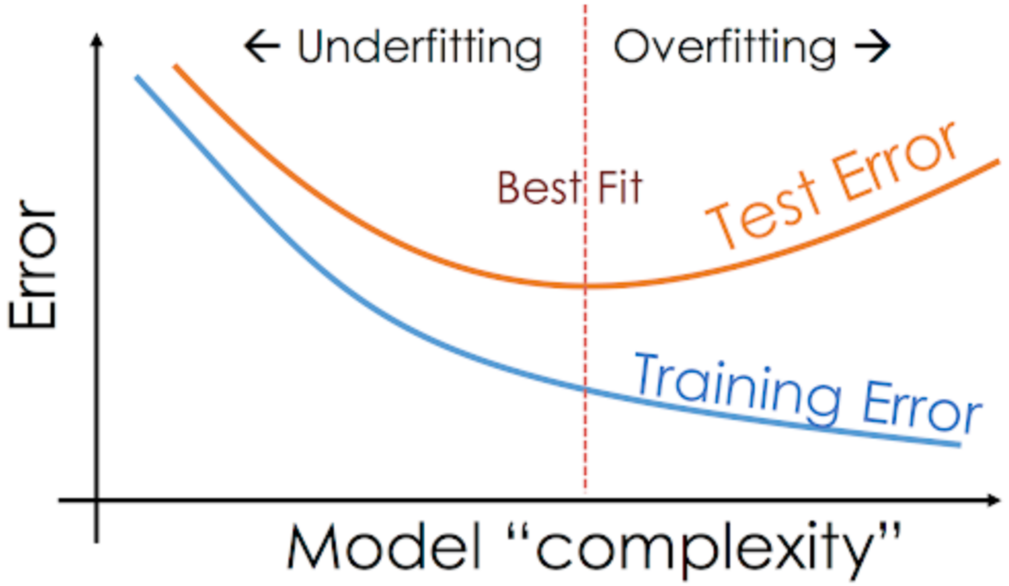


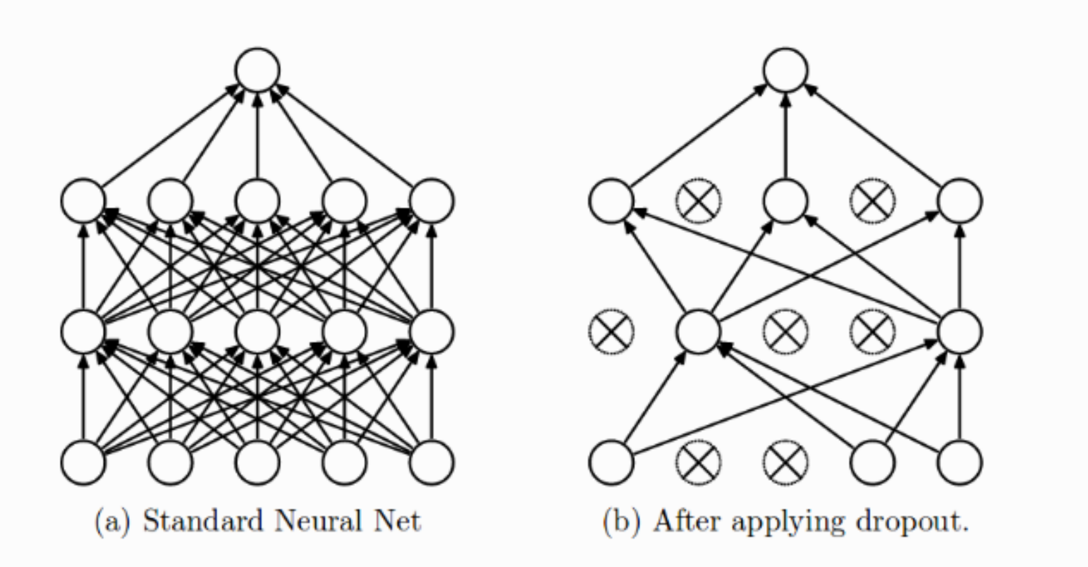
https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-9/v-4/

Overfitting

Model-level solution

- Reduce model complexity by introducing regularization techniques.
    - L2 normalization, dropout, ensembles, label smoothing, etc.





https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

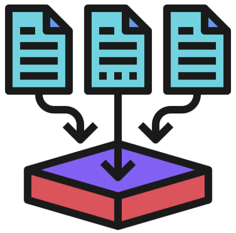
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural ne tworks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Overfitting

Data-level solution

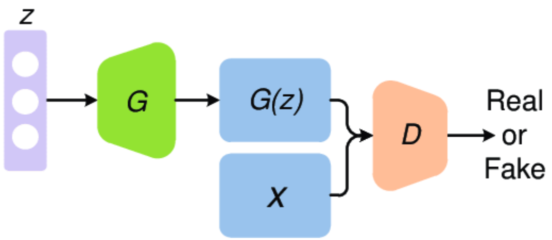
- Increase the size of training data to better approximate the true data distribution.

1. Collect more



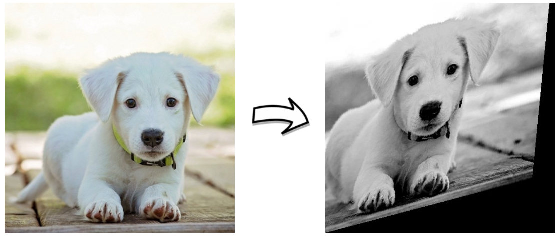
- ❑ Expensive
- ❑ EXPENSIVE.
- ❑ *E.X.P.E.N.S.I.V.E.*

2. Synthesize



- ❑ Complicated
- ❑ Little variation
- ❑ Mode collapse

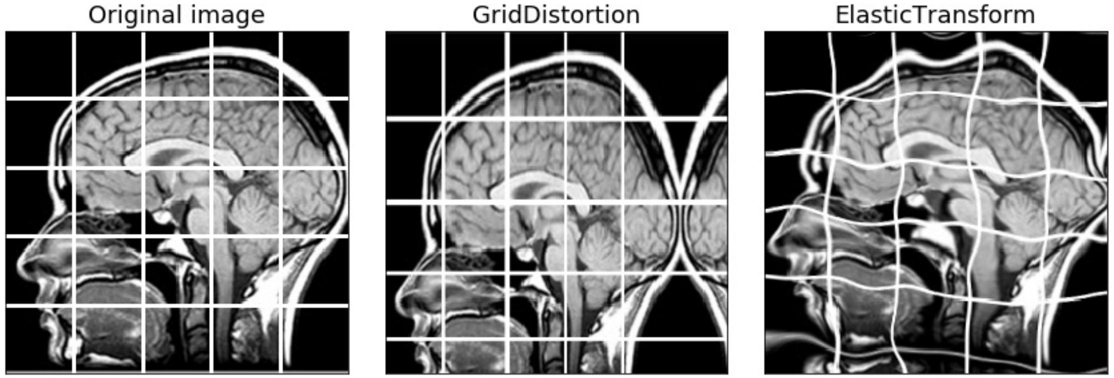
3. Augment

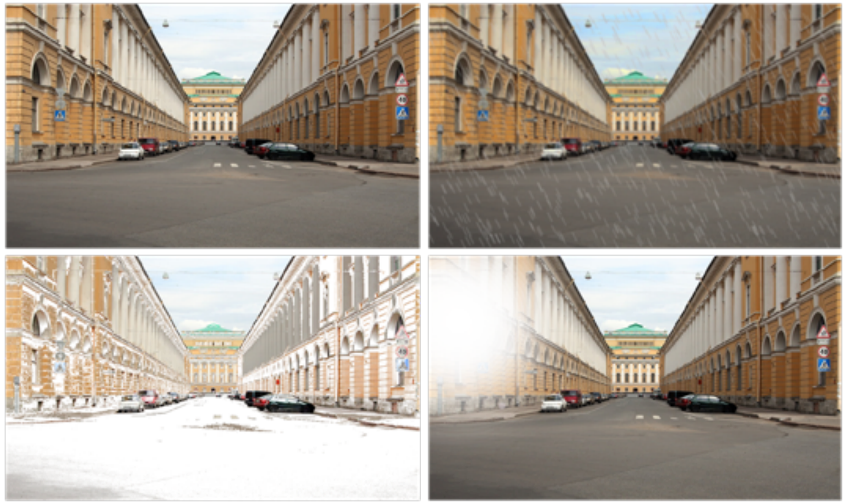


- ❑ Simple
- ❑ Well-studied
- ❑ Which is optimal?

Data Augmentation

Examples with images





https://github.com/albumentations-team/albumentations

Data Augmentation

How much?

<u>Semantically</u> <u>Invariant</u> Transformation

Adverb. 의미상, 의미론적으로 Adjective. 변함없는, 변치 않는

=

Transformations should preserve class labels.

Data Augmentation

How much?

A **hamster**

augmentation

Crop
Rotate
Contrast
Invert
Grayscale
…

is still a **hamster**.

https://github.com/aleju/imgaug

Data Augmentation

How much?

A **Hyungu**

augmentation

Crop
Rotate
Contrast
Invert
Grayscale
…

is still a **Hyungu**?

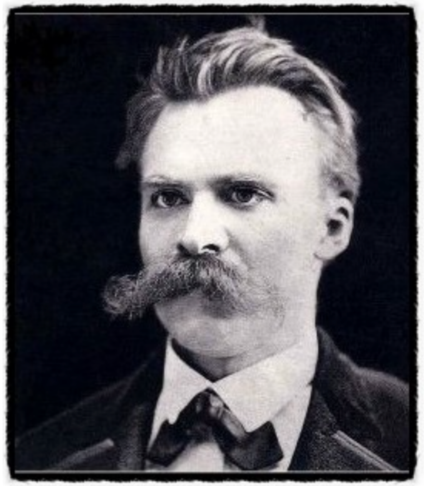
https://frozen.disney.com/olaf

Data Augmentation

How much?

What doesn’t kill you makes you stronger.

널 죽이지 못하는 것은 너를 더 강하게 만들 뿐이다.



Friedrich Nietzsche (1844 ~ 1900)

Kelly Clarkson (1982 ~)

https://www.youtube.com/watch?list=PL805Ei3KqUWsr2i-D9ZROH_w7M93GLkle&v=Xn676-fLq7I

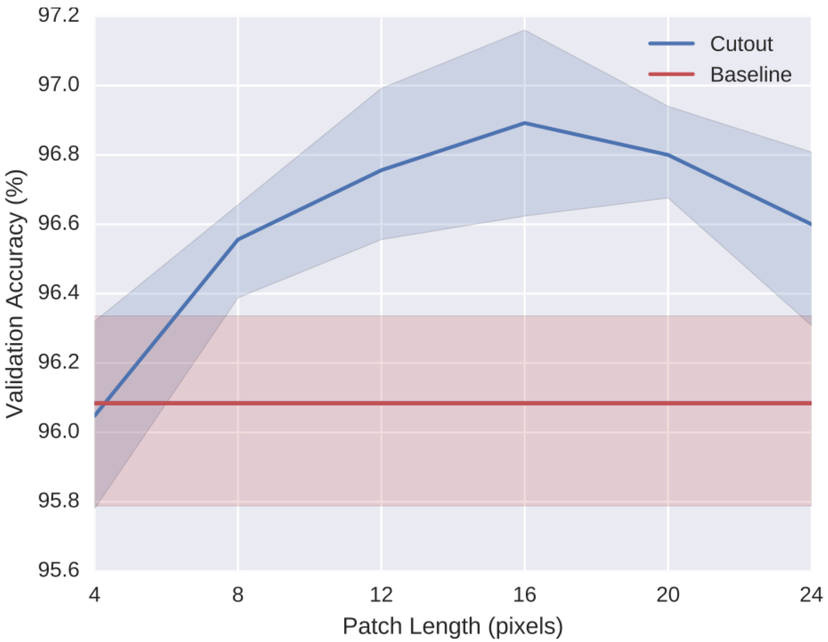
Data Augmentation for Computer Vision

Cutout

DeVries and Taylor, 2017

- Randomly replace square patches with noise.

CIFAR-10 classification

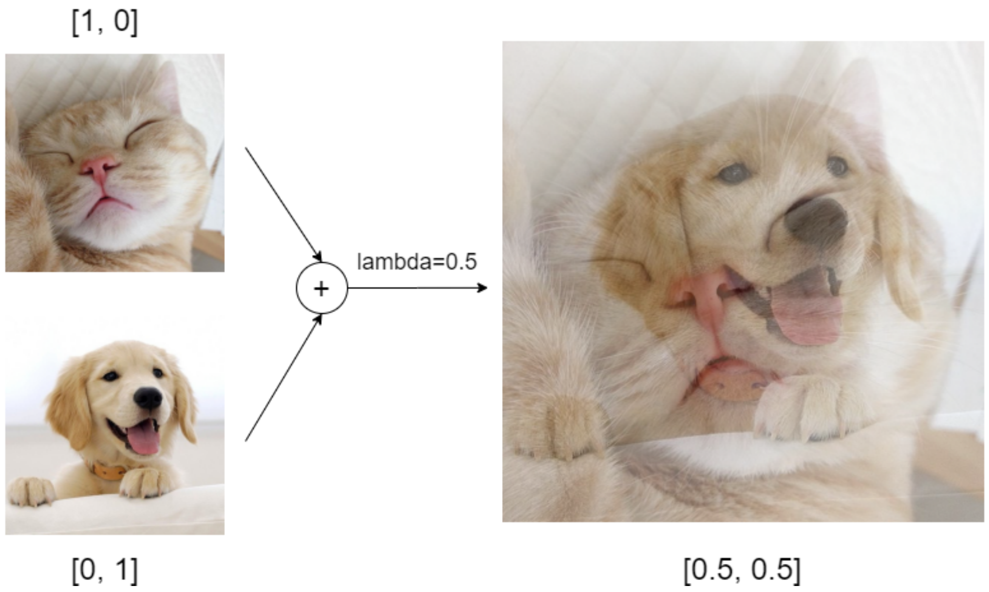


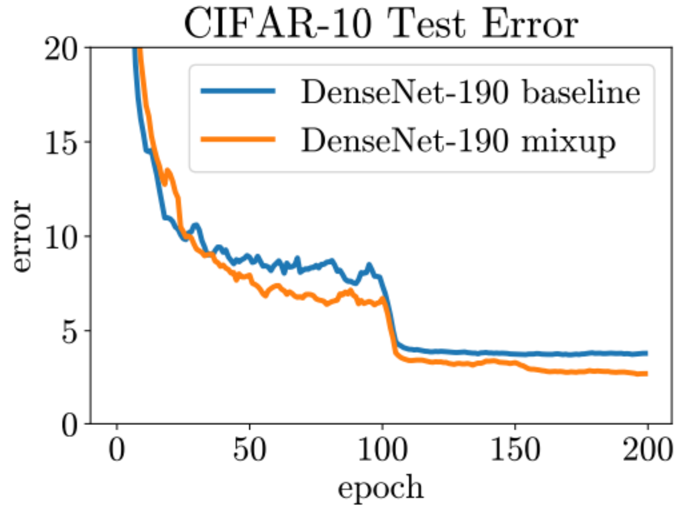
DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Mixup

Zhang et al., 2018

- Convex combination of pairs of images and their class labels.





(b) Test error evolution for the best ERM and *mixup* models.

https://medium.com/@wolframalphav1.0/easy-way-to-improve-image-classifier-performance-part-1-mixup-augmentation-with-codes-33288db92de5

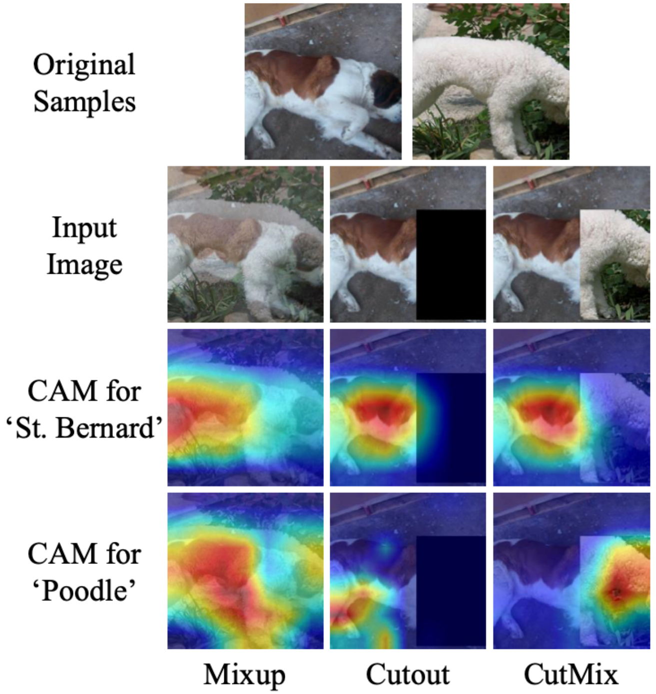
Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

CutMix

Yun et al., 2019

- Patches are cut and pasted among training images.
- Class labels are also mixed proportionally to the area of the patches.

	ResNet-50	Mixup [48]	Cutout [3]	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	**78.6 (+2.3)**
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	**47.3 (+1.0)**
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	**76.7 (+1.1)**

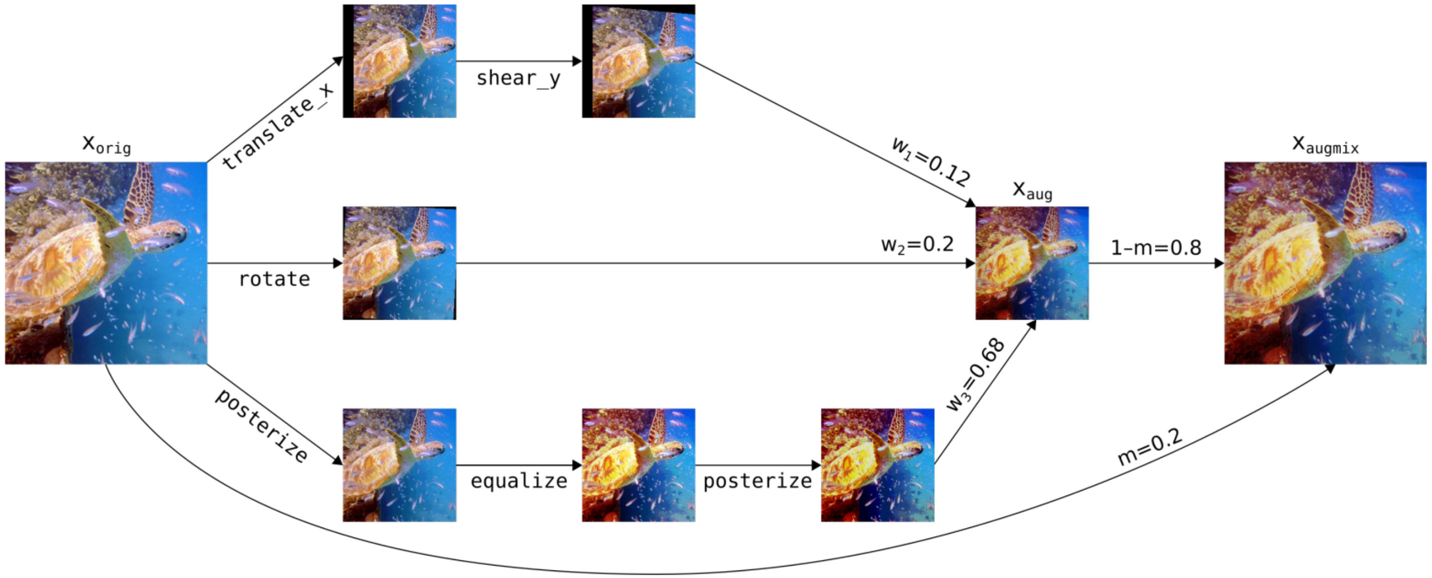


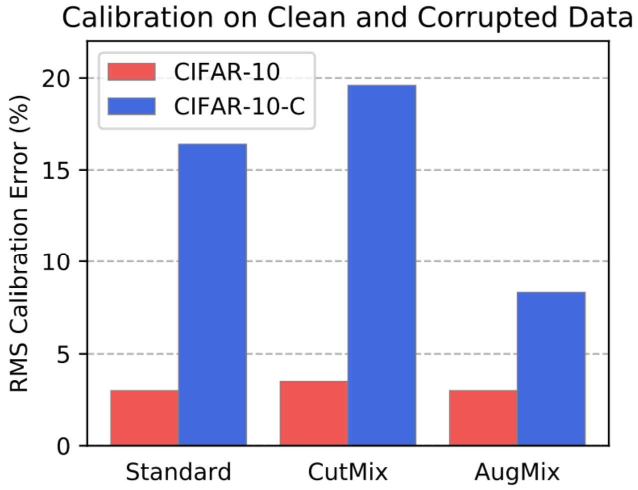
Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6023-6032).

AugMix

Hendrycks et al., 2019

- Create multiple augmented images and mix them.
- Augmentation operations and the mixing weights are randomly sampled.
- Improves noise robustness and uncertainty estimates.



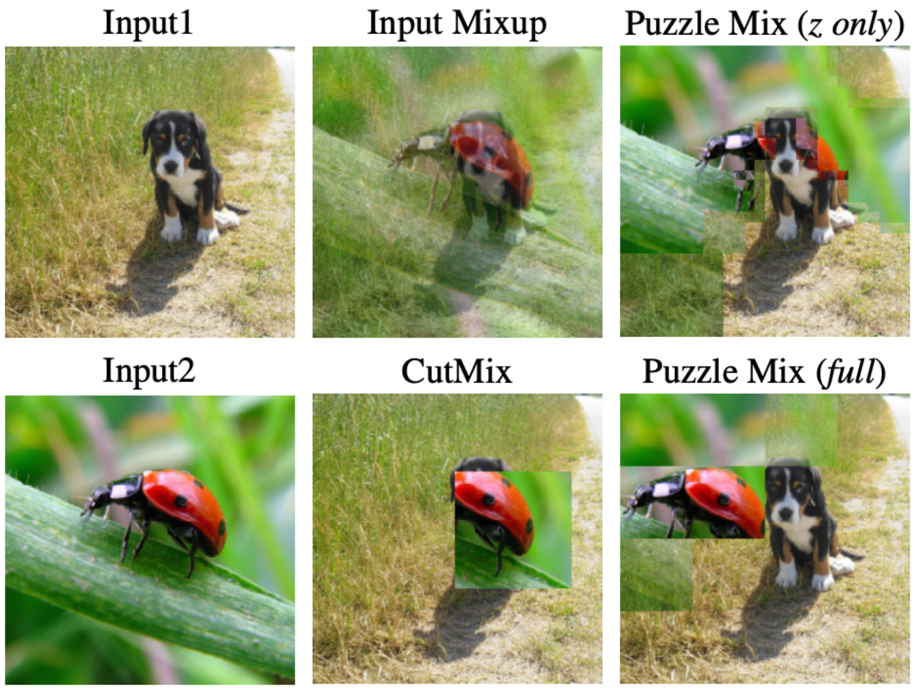


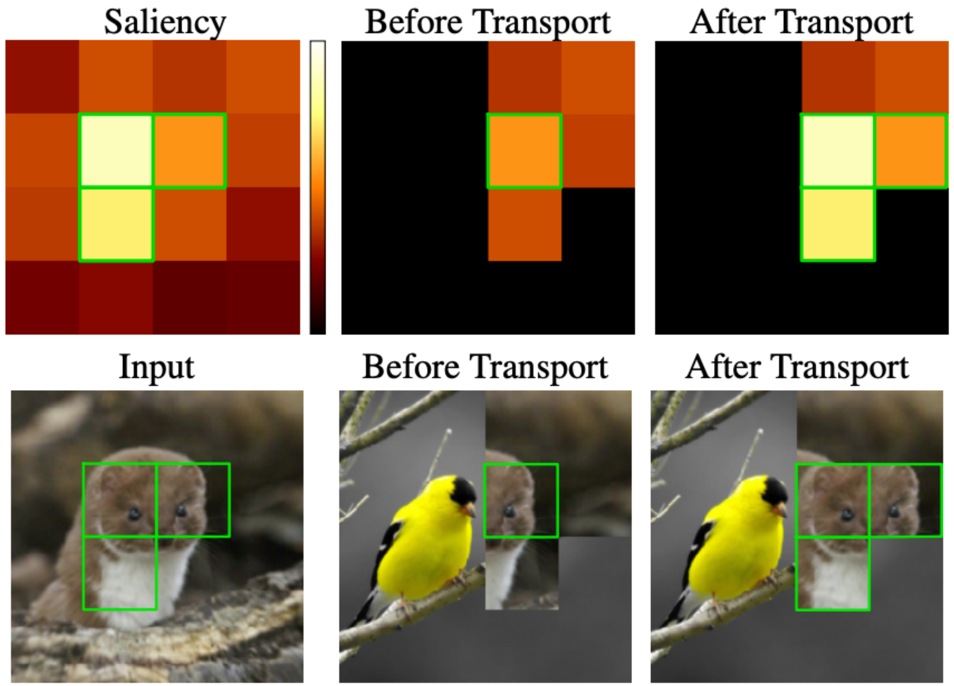
Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2019). Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781.

Puzzle Mix

Kim et al., 2020

- Utilize the regional saliency information of natural images.
- Solve an binary transport problem to find the optimal move that maximizes saliency.





Kim, J. H., Choo, W., & Song, H. O. (2020, November). Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In International Confe rence on Machine Learning (pp. 5275-5285). PMLR.

Which is optimal?

Can we find better ones?

Why not automate the search process?

AutoAugment (AA)

Cubuk et al., 2019

- Search for the best policy of augmentations using ***reinforcement learning (PPO)***.
- Extremely slow; takes up to 5,000 hours on CIFAR-10, and 15,000 hours on ImageNet.

Operation Name	Description	Range of magnitudes
ShearX(Y)	Shear the image along the horizontal (vertical) axis with rate *magnitude*.	[-0.3,0.3]
TranslateX(Y)	Translate the image in the horizontal (vertical) direction by *magnitude* number of pixels.	[-150,150]
Rotate	Rotate the image *magnitude* degrees.	[-30,30]
AutoContrast	Maximize the the image contrast, by making the darkest pixel black and lightest pixel white.	
Invert	Invert the pixels of the image.	
Equalize	Equalize the image histogram.	
Solarize	Invert all pixels above a threshold value of *magnitude*.	[0,256]
Posterize	Reduce the number of bits for each pixel to *magnitude* bits.	[4,8]
Contrast	Control the contrast of the image. A *magnitude*=0 gives a gray image, whereas *magnitude*=1 gives the original image.	[0.1,1.9]
Color	Adjust the color balance of the image, in a manner similar to the controls on a colour TV set. A *magnitude*=0 gives a black & white image, whereas *magnitude*=1 gives the original image.	[0.1,1.9]
Brightness	Adjust the brightness of the image. A *magnitude*=0 gives a black image, whereas *magnitude*=1 gives the original image.	[0.1,1.9]
Sharpness	Adjust the sharpness of the image. A *magnitude*=0 gives a blurred image, whereas *magnitude*=1 gives the original image.	[0.1,1.9]
Cutout [25, 72]	Set a random square patch of side-length *magnitude* pixels to gray.	[0,60]
Sample Pairing [50, 73]	Linearly add the image with another image (selected at random from the same mini-batch) with weight *magnitude*, without changing the label.	[0, 0.4]

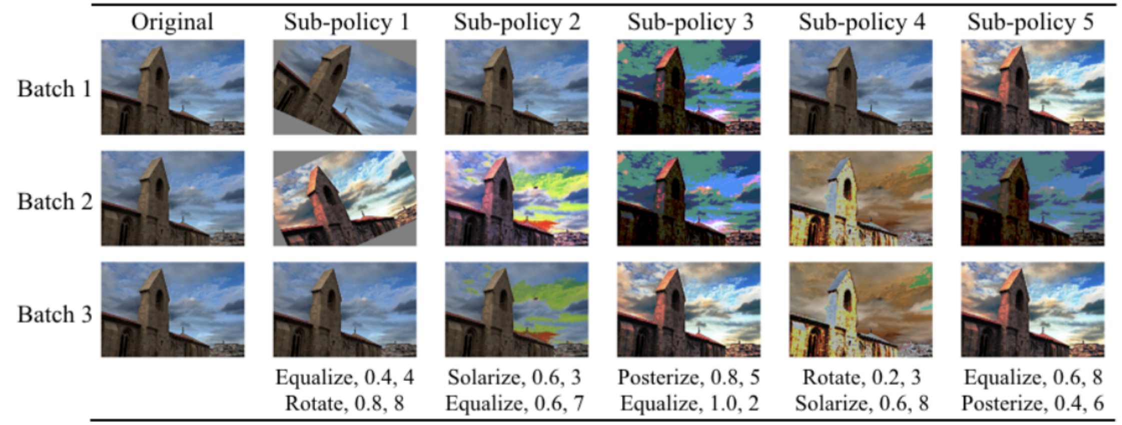


Figure 3. One of the successful policies on ImageNet. As described in the text, most of the policies found on ImageNet used color-based transformations.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 113-123).

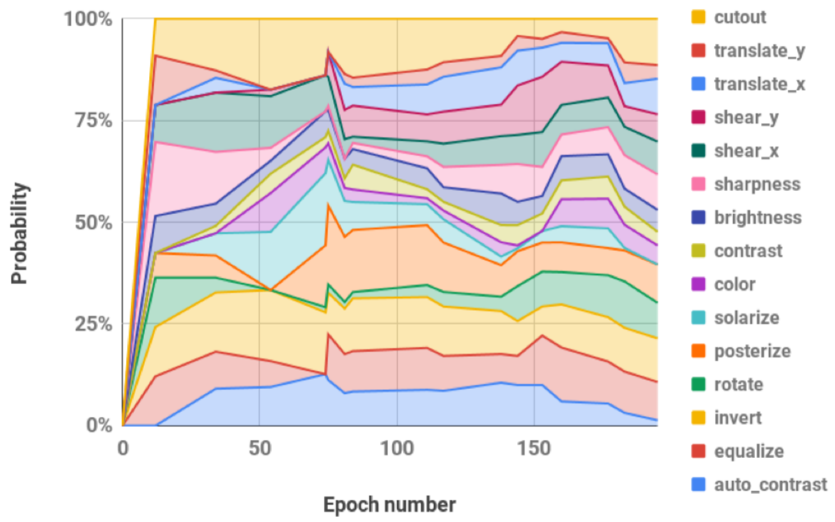
Population-Based Augmentation (PBA)

Ho et al., 2019

- Uses ***population-based training*** to speed up the learning process.
- Learns an augmentation schedule instead of a fixed augmentation policy.

Dataset	Value	Previous Best	AA	PBA
CIFAR-10	GPU Hours	-	5000	5
	Test Error	2.1	1.48	1.46
CIFAR-100	GPU Hours	-	0*	0*
	Test Error	12.2	10.7	10.9
SVHN	GPU Hours	-	1000	1
	Test Error	1.3	1.0	1.1

Table 1. Comparison of pre-computation costs and test set error (%) between this paper, AutoAugment (AA), and the previous best published results.



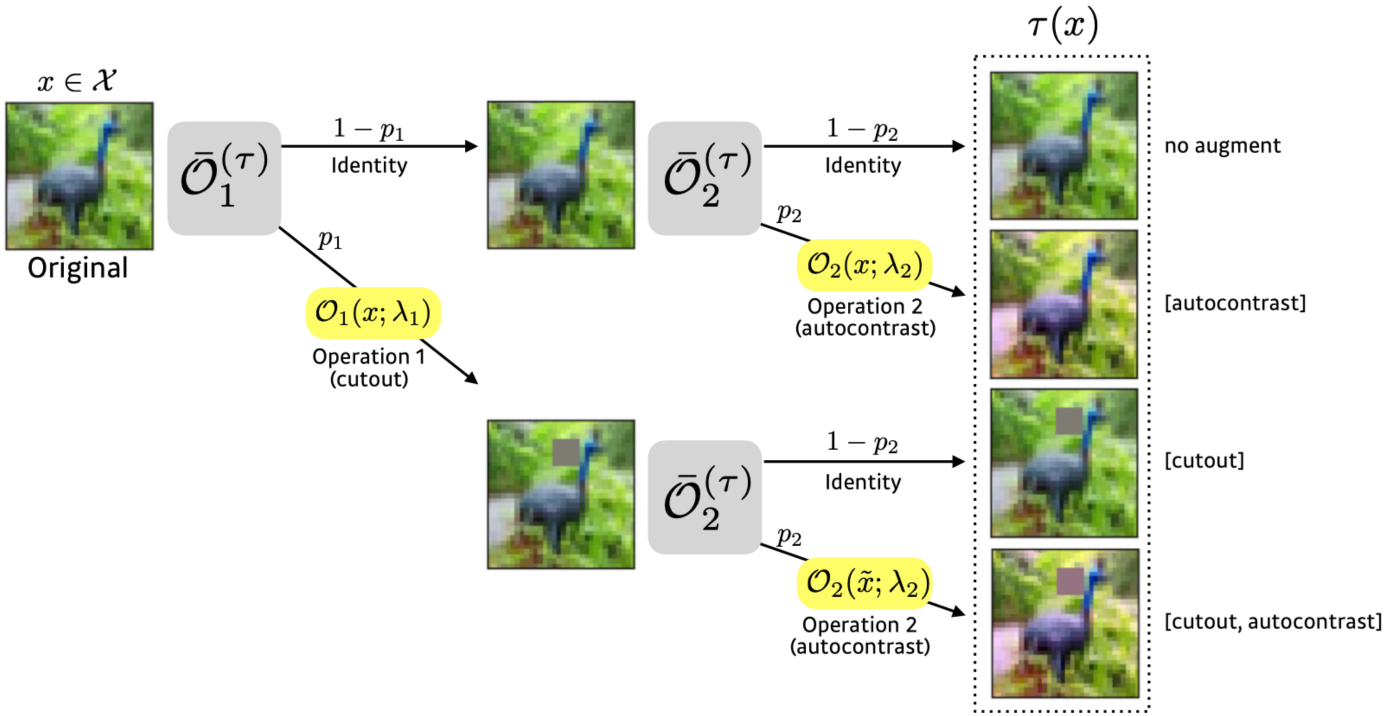
(b) Normalized plot of operation probability parameters over time. The distribution flattens out towards the end of training.

Ho, D., Liang, E., Chen, X., Stoica, I., & Abbeel, P. (2019, May). Population based augmentation: Efficient learning of augmentation policy schedul es. In International Conference on Machine Learning (pp. 2731-2741). PMLR.

Fast AutoAugment (Fast AA)

Lim et al., 2019

- Use ***Bayesian optimization*** techniques to speed up the search process.
- Reduces the computational cost to 3.5 hours on CIFAR-10, and 450 hours on ImageNet.



Dataset	AutoAugment [3]	Fast AutoAugment
CIFAR-10	5000	3.5
SVHN	1000	1.5
ImageNet	15000	450

Table 1: GPU hours comparison of Fast AutoAugment with AutoAugment.

Model	Baseline	AutoAugment [3]	Fast AutoAugment
ResNet-50	23.7 / 6.9	**22.4 / 6.2**	**22.4** / 6.3
ResNet-200	21.5 / 5.8	20.00 / 5.0	**19.4 / 4.7**

Table 5: Validation set Top-1 / Top-5 error rate (%) on ImageNet.

Lim, S., Kim, I., Kim, T., Kim, C., & Kim, S. (2019). Fast autoaugment. In Advances in Neural Information Processing Systems (pp. 6665-6675).

RandAugment (RA)

Cubuk et al., 2020

- identity
- rotate
- posterize
- sharpness
- translate-x
- autoContrast
- solarize
- contrast
- shear-x
- translate-y
- equalize
- color
- brightness
- shear-y

- ***Randomly sample*** a subset from a predefined set of 14 image transforms.
- Sequentially apply them with random distortion magnitudes.

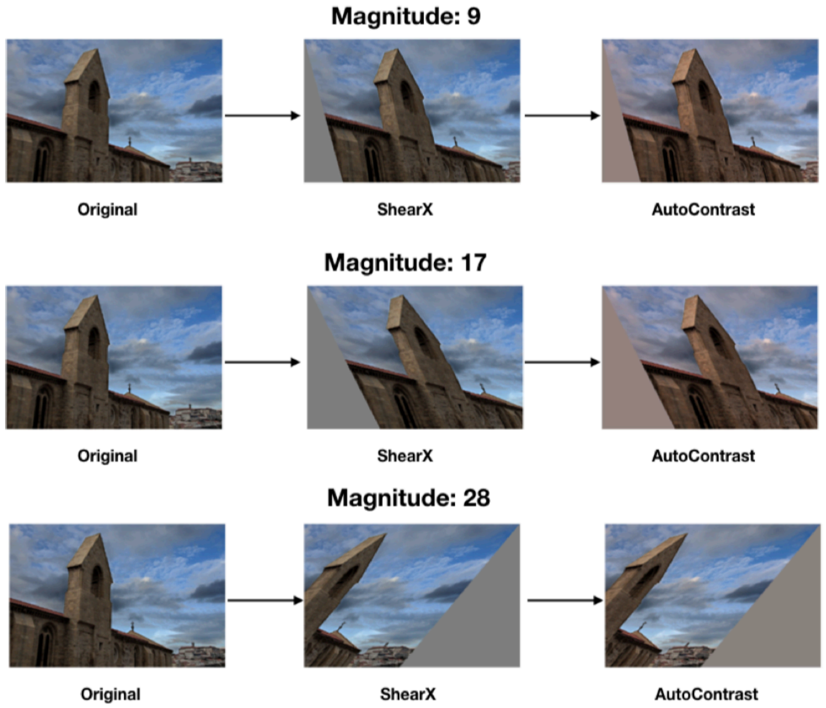


Figure 1. **Example images augmented by RandAugment.**

	baseline	PBA	Fast AA	AA	RA
CIFAR-10					
Wide-ResNet-28-2	94.9	-	-	**95.9**	95.8
Wide-ResNet-28-10	96.1	**97.4**	97.3	**97.4**	97.3
Shake-Shake	97.1	**98.0**	**98.0**	**98.0**	**98.0**
PyramidNet	97.3	**98.5**	98.3	**98.5**	**98.5**
CIFAR-100					
Wide-ResNet-28-2	75.4	-	-	**78.5**	78.3
Wide-ResNet-28-10	81.2	**83.3**	82.7	82.9	**83.3**
SVHN (core set)					
Wide-ResNet-28-2	96.7	-	-	98.0	**98.3**
Wide-ResNet-28-10	96.9	-	-	98.1	**98.3**
SVHN					
Wide-ResNet-28-2	98.2	-	-	**98.7**	**98.7**
Wide-ResNet-28-10	98.5	98.9	98.8	98.9	**99.0**

Table 2. **Test accuracy (%) on CIFAR-10, CIFAR-100, SVHN and SVHN core set.**

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 702-703).

Data Augmentation for Natural Language Processing

Thesaurus-based Substitution

Zhang et al., 2015

- Replace a random word with its synonym using a Thesaurus.

https://amitness.com/2020/05/data-augmentation-for-nlp/

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing s ystems, 28, 649-657.

Word Embedding-based Substitution

Jiao et al., 2019

- Use pre-trained word embeddings such as Word2Vec, GloVe, FastText, etc.
- Find the nearest neighbor words and substitute.

https://amitness.com/2020/05/data-augmentation-for-nlp/

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.

Masked Language Model

Garg et al., 2020

- Use pre-trained masked language models such as BERT, ROBERTa, and ALBERT.
- Mask out some words & see what the model predicts.

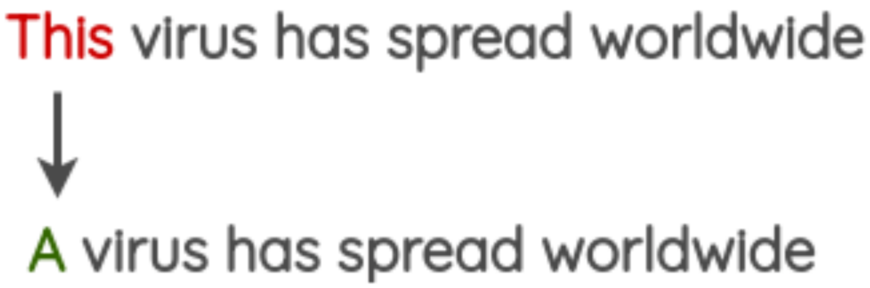
https://amitness.com/2020/05/data-augmentation-for-nlp/

Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based Adversarial Examples for Text Classification. arXiv preprint arXiv:2004.01970.

TF-IDF-based Replacement

Xie et al., 2019

- Words with low TF-IDF scores are *uninformative*.
- Replacing those words will *not* affect the original semantic information.



https://amitness.com/2020/05/data-augmentation-for-nlp/

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848.

Back Translation

Xie et al., 2019

- Translate into another language.
- Translate back to original language.

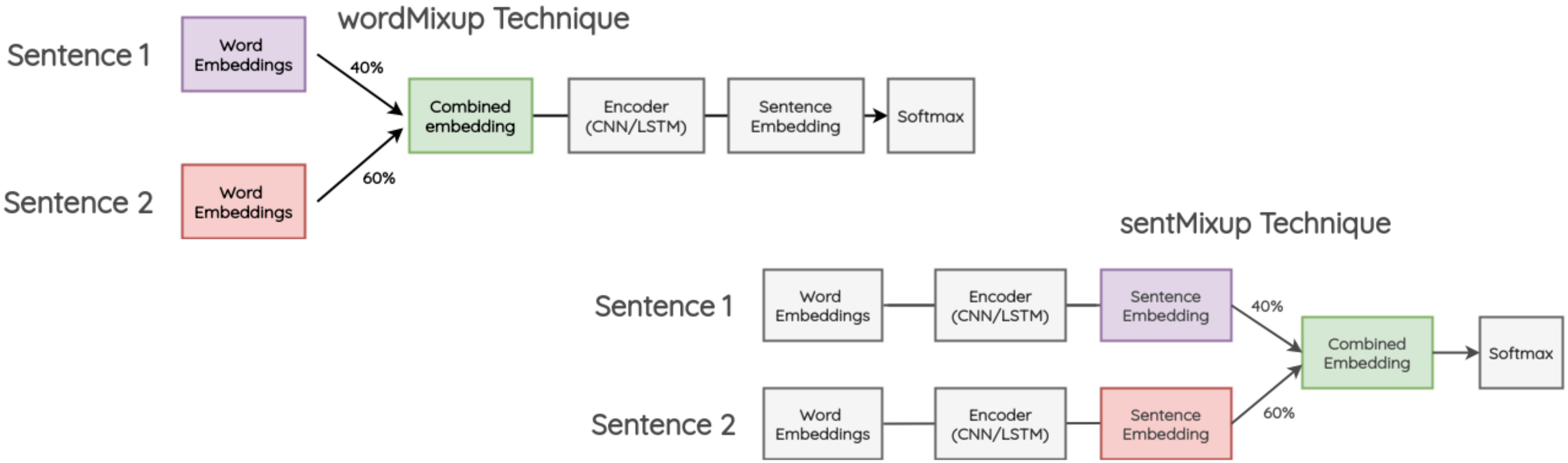
https://amitness.com/2020/05/data-augmentation-for-nlp/

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848.

Word/Sent Mixup

Xie et al., 2019

- Use Mixup on word embedding features
- Use Mixup on sentence embedding features



https://amitness.com/2020/05/data-augmentation-for-nlp/

Guo, H., Mao, Y., & Zhang, R. (2019). Augmenting data with mixup for sentence classification: An empirical study. arXiv preprint arXiv:1905.08941.

Conclusions

- Data augmentation reduces overfitting on the training data.
- Various techniques have been developed for CV & NLP.

- Subfields of machine learning that leverage data augmentation
  - ❑Semi-supervised learning w/ consistency regularization
  - ❑Self-supervised contrastive learning

- Data augmentation for other data domains
  - ❑Audio
  - ❑Graphs
  - ❑Structured tabular data

Thank you.