DSME 6635: Artificial Intelligence for Business Research

Deep-Learning-Based Image Classification

Renyu (Philip) Zhang

1

Visual vs. Language

- Visual data or language, which one is more informative?
- Is a picture worth 1,000 words?

References: https://lexfridman.com/yann-lecun-3-transcript
https://www.linkedin.com/posts/yann-lecun_parm-prmshra-on-x-activity-7172266619103080448-iqvP/

Yann LeCun · 3rd+
VP & Chief AI Scientist at Meta
2w · Edited ·

✓ Following ···

* Language is low bandwidth: less than 12 bytes/second. A person can read 270 words/minutes, or 4.5 words/second, which is 12 bytes/s (assuming 2 bytes per token and 0.75 words per token). A modern LLM is typically trained with 1x10^13 two-byte tokens, which is 2x10^13 bytes. This would take about 100,000 years for a person to read (at 12 hours a day).

* Vision is much higher bandwidth: about 20MB/s. Each of the two optical nerves has 1 million nerve fibers, each carrying about 10 bytes per second. A 4 year-old child has been awake a total 16,000 hours, which translates into 1x10^15 bytes.

In other words:
- The data bandwidth of visual perception is roughly 1.6 million times higher than the data bandwidth of written (or spoken) language.
- In a mere 4 years, a child has seen 50 times more data than the biggest LLMs trained on all the text publicly available on the internet.

This tells us three things:
1. Yes, text is redundant, and visual signals in the optical nerves are even more redundant (despite being 100x compressed versions of the photoreceptor outputs in the retina). But redundancy in data is *precisely* what we need for Self-Supervised Learning to capture the structure of the data. The more redundancy, the better for SSL.
2. Most of human knowledge (and almost all of animal knowledge) comes from our sensory experience of the physical world. Language is the icing on the cake. We need the cake to support the icing.
3. There is *absolutely no way in hell* we will ever reach human-level AI without getting machines to learn from high-bandwidth sensory inputs, such as vision.

Yes, humans can get smart without vision, even pretty smart without vision and audition. But not without touch. Touch is pretty high bandwidth, too.

2

2

Agenda

- Image Classification
- Image Classification Applications in Business/Econ Research

3

3

Image Processing and Computer Vision

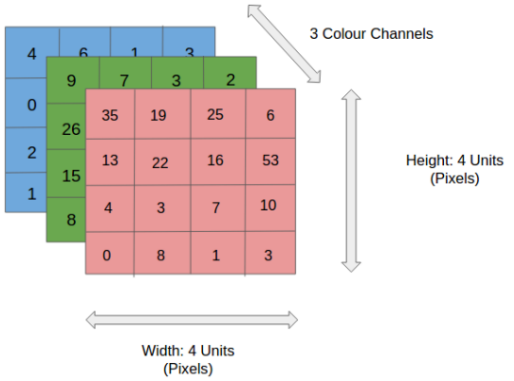
- **Image Processing**: The use of a digital computer to process digital images through an algorithm.
- **Computer Vision**: An interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images and videos.

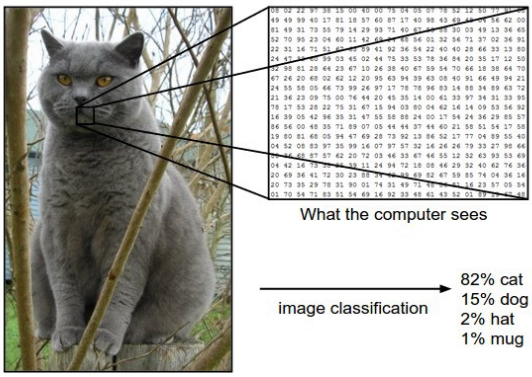
4

4

Image Representation

- Reference: https://cs231n.github.io/classification/
- Again, there's a representation problem and a subsequent classification problem.
- Compared with language, there's a very natural and clear physical representation, but there's no inherent logic that supports self-supervised learning.



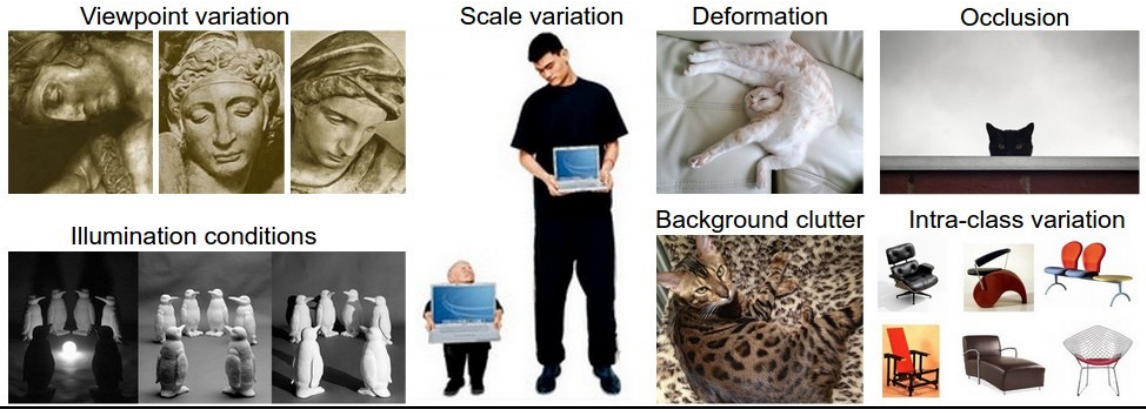


5

5

Image Classification

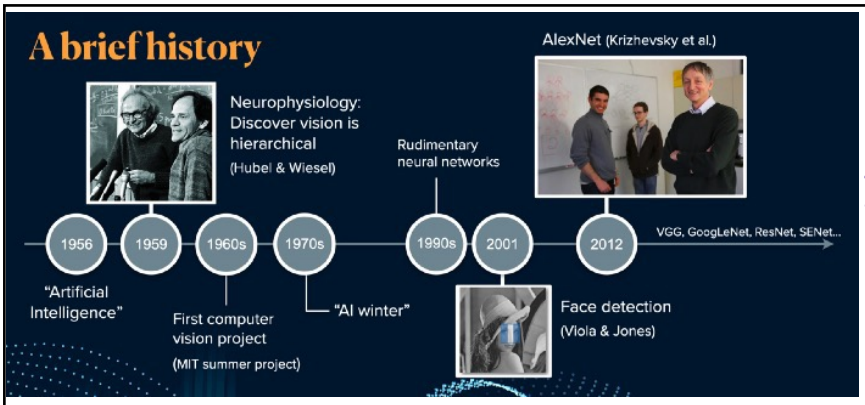
- Reference: https://cs231n.github.io/classification/
- Unlike NLP, there is one most important task of image processing: Object classification.
- Image classification algorithms are tested on ImageNet (https://www.image-net.org/).
- Image recognition/classification is challenging in general.

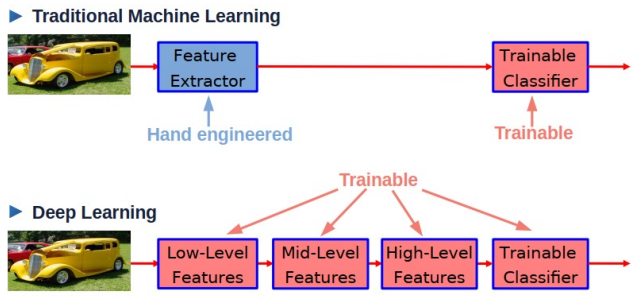


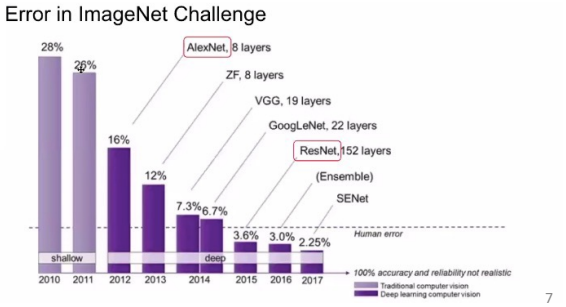
6

6

Image Classification: A Brief History

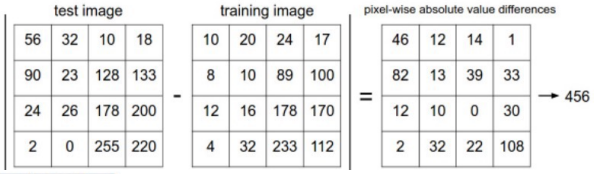






Ancient Techniques

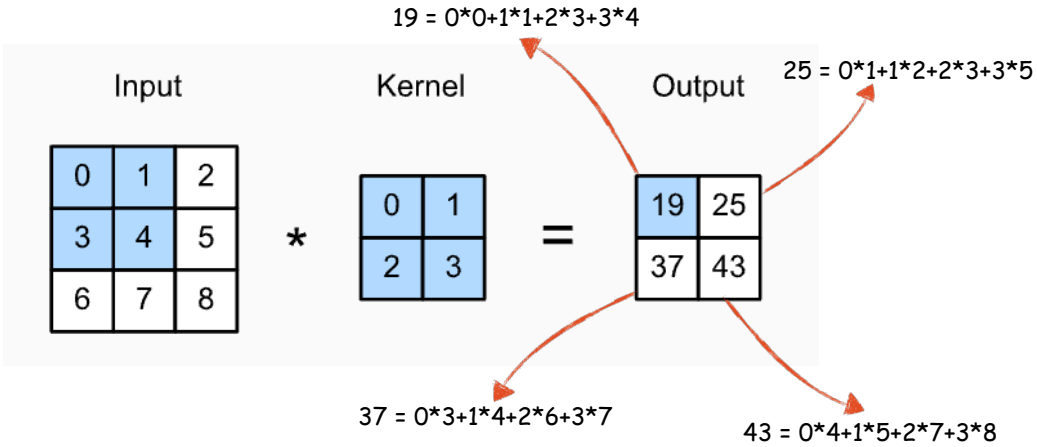
- Reference: https://cs231n.github.io/classification/
- A popular toy data set: CIFAR-10 (https://www.cs.toronto.edu/~kriz/cifar.html)
- Non-DL methods involve nearest neighbors and SVM.



Left: Example images from the CIFAR-10 dataset. Right: first column shows a few test images and next to each we show the top 10 nearest neighbors in the training set according to pixel-wise difference.

Convolution Neural Network (CNN)

- Reference: https://cs231n.github.io/convolutional-networks/
- The convolution layer is a transformation pixel by pixel, done by applying to an image some transformation defined by a set of weights, known as filters/kernels.



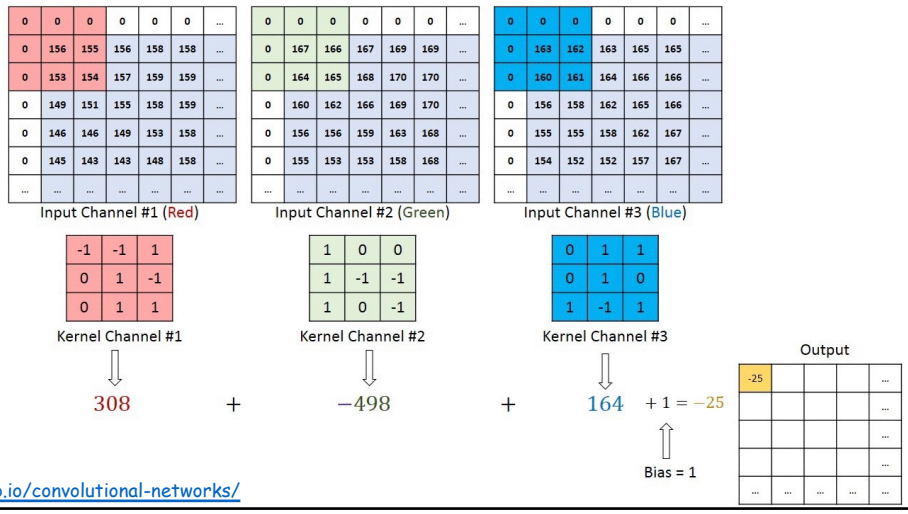
9

9

Convolution Neural Network (CNN)

- Some additional parameters for convolution:
  - Padding: Adding values around the original image so that each pixel is covered the same times.
  - Stride: How many pixels should each step skip; it is used not to double count pixels too much.

$$o = \frac{i + 2p - k}{s} + 1$$



Reference: https://cs231n.github.io/convolutional-networks/

10

10

Why Does CNN Work?

- Geometrically, it allows you to emphasize on certain features that are rotational or positional invariant.
- CNNs are extremely prone to overfitting, so they are easy to estimate with reasonable amount of information.

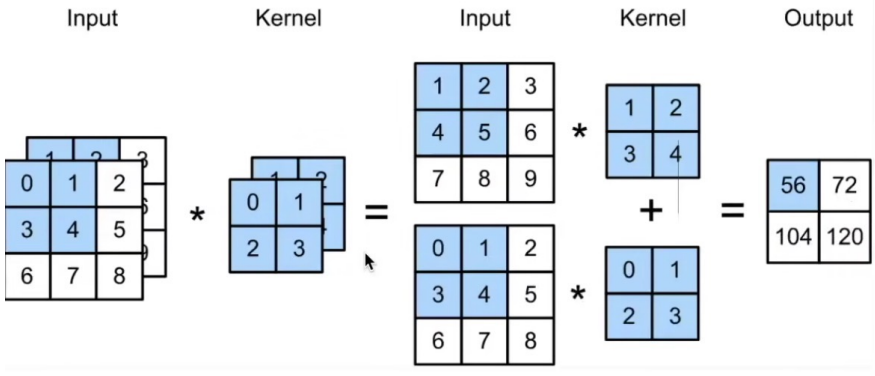
Reference: https://cs231n.github.io/convolutional-networks/

11

11

Multi-Channel CNN

- One more dimension: The number of input channel.



$$(1\times 1+2\times 2+4\times 3+5\times 4)$$
$$+(0\times 0+1\times 1+3\times 2+4\times 3)=56$$

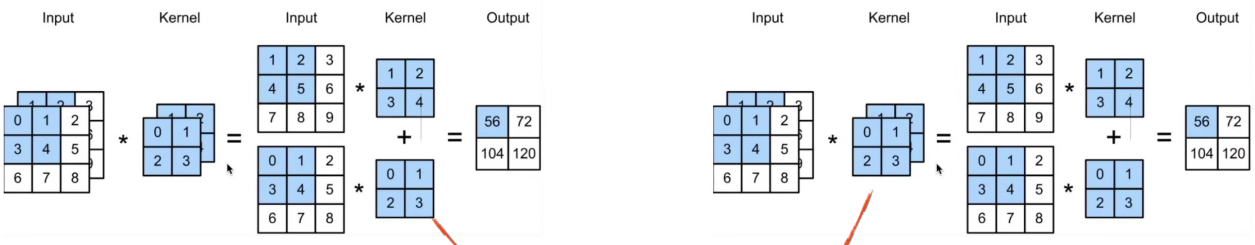
Reference: https://www.d2l.ai/chapter_convolutional-neural-networks/channels.html

12

12

Multi-Channel CNN

- Last parameter: The number of output dimension.



Output Channel 1

Output Channel 2

In general, the kernels of different output channels are different.

Reference: https://www.d2l.ai/chapter_convolutional-neural-networks/channels.html

13

13

Putting Everything Together

- Dimensions:
  - Input X: C_i * N_h * N_w
  - Kernel W: C_o * C_i * K_h * K_w
  - Bias B: C_o * M_h * M_w
  - Output Y: C_o * M_h * M_w

$$o = \frac{i + 2p - k}{s} + 1$$

$$\mathbf{Y} = \mathbf{X} \star \mathbf{W} + \mathbf{B}$$

Convolution

- Computational complexity:
  - O(C_i * C_o * K_h * K_w * M_h * M_w)
  - Assume: C_i = C_o = 100, K_h= K_w = 5, M_h = M_w = 65
  - ~1 GFLOPS for one CNN layer
  - Assume: 10 CNN layers, 1 million images, which means 10 PFLOPS in total.
  - If we use CPU, the time for one pass: 10 * 2 PFLOPS / 0.15 TFLOPS ~ 18 hours.
  - If we use GPU, say 3090, the time for one pass: 10 * 2 PFLOPS / 40 TFLOPS ~ 8.3 mins.
  - Exactly what happened to AlexNet.

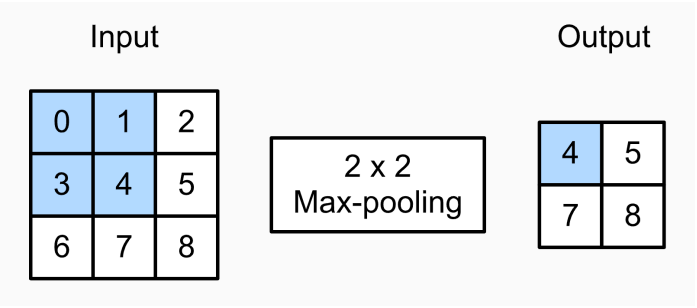
Reference: https://cs231n.github.io/convolutional-networks/

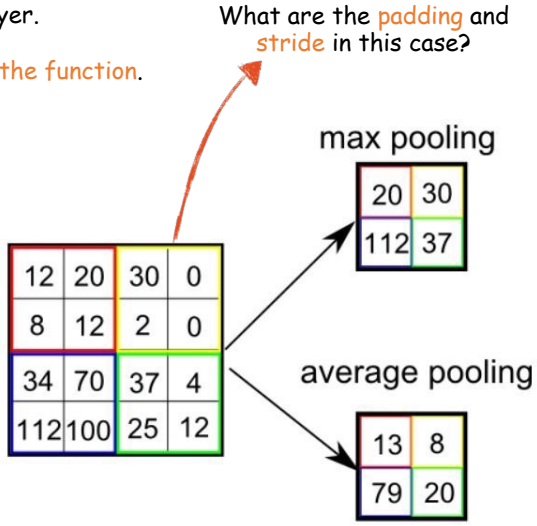
14

14

Pooling

- You can do (max or average) pooling after the convolution layer.
- This is another way to reduce dimensionality and regularize the function.
- You may also do padding and stride.





Reference: https://www.d2l.ai/chapter_convolutional-neural-networks/pooling.html

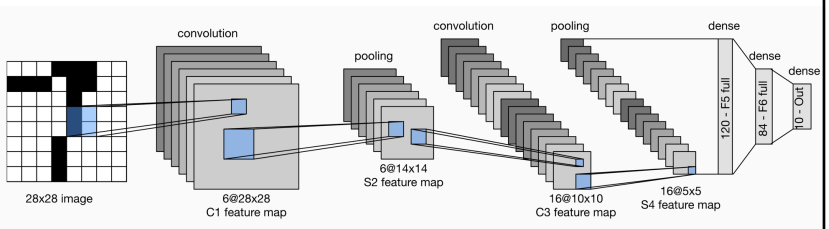
15

15

LeNet (1998)

Gradient-based learning applied to document recognition
Y LeCun, L Bottou, Y Bengio... - Proceedings of the ..., 1998 - ieeexplore.ieee.org
... gradientbased learning technique. Given an appropriate network architecture, gradient-based learning ... This paper reviews various methods applied to handwritten character recognition ...
☆ 保存 99 引用 被引用次数：63071 相关文章 所有 40 个版本 Web of Science: 28544 »

- The first demonstration of CNN in computer vision (handwriting recognition).
- Modified national institute of standards and technology (MNIST) Dataset: 60,000 images (50,000 training + 10,000 testing) of handwritten digits.



Reference: https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

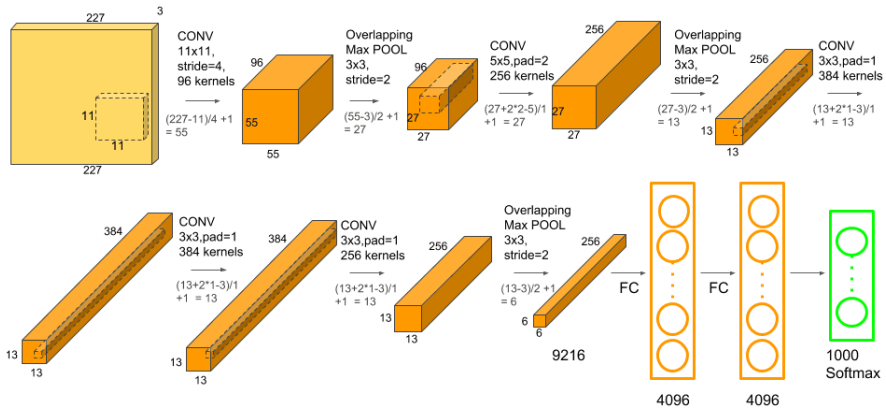
16

16

AlexNet (2012)

Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever… - Advances in neural …, 2012 - proceedings.neurips.cc
… We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test …
☆ 保存 99 引用 被引用次数：127107 相关文章 所有 102 个版本 »

- AlexNet demonstrated a substantial improvement in ImageNet classification error over the best non-DL method by a wide margin (25.8% → 16.4%, human error = 5.1%).



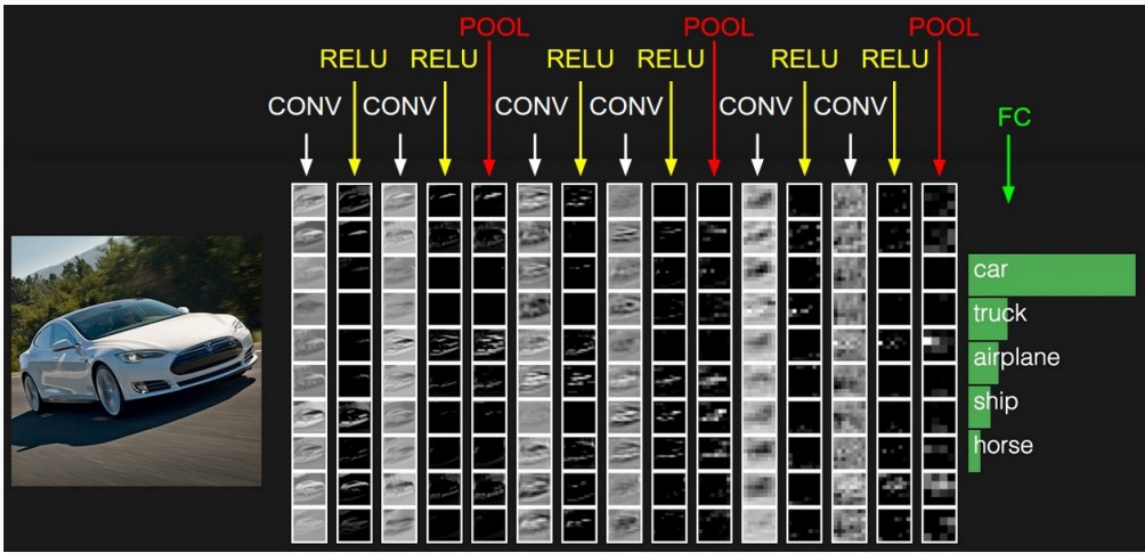
Reference: http://cs231n.stanford.edu/slides/2023/lecture_6.pdf

17

17

AlexNet (2012)

Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever… - Advances in neural …, 2012 - proceedings.neurips.cc
… We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test …
☆ 保存 99 引用 被引用次数：127107 相关文章 所有 102 个版本 »



Reference: http://cs231n.stanford.edu/slides/2023/lecture_6.pdf

18

18

AlexNet vs. LeNet

- Nothing new, just deeper, larger, and more data.
- LeNet: 0.1M - 0.5M parameters
- AlexNet: 40M - 60M parameters
- LeNet: 4M FLOPS for one image
- AlexNet: 1G FLOPS for one image

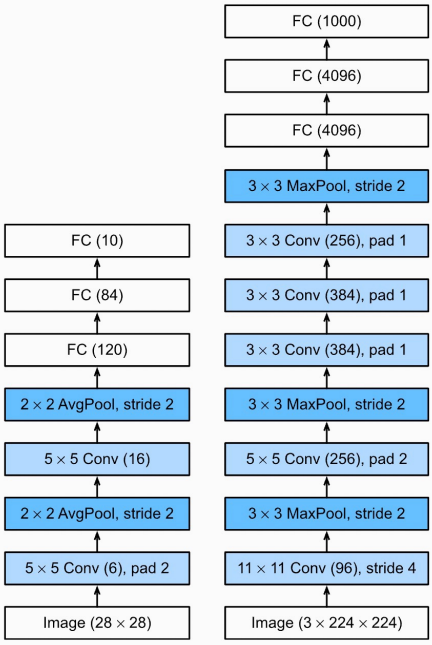


Fig. 8.1.2 From LeNet (left) to AlexNet (right).

Reference: https://www.d2l.ai/chapter_convolutional-modern/alexnet.html

19

19

After AlexNet

Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever… - Advances in neural …, 2012 - proceedings.neurips.cc
… We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test …
☆ 保存 99 引用 被引用次数：127107 相关文章 所有 102 个版本 »

Deep residual learning for image recognition
K He, X Zhang, S Ren, J Sun - Proceedings of the IEEE …, 2016 - openaccess.thecvf.com
… Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. …
★ 保存 99 引用 被引用次数：209436 相关文章 所有 73 个版本 »

LeNet (1998) — AlexNet (2012) — VGG (2014) — ResNet (2016) — SENet (2018)

8 layers
Error = 16.4%

19 layers
Error = 7.3%

152 layers
Error = 3.6%

152 layers
Error = 2.3%

Gradient-based learning applied to document recognition
Y LeCun, L Bottou, Y Bengio… - Proceedings of the …, 1998 - ieeexplore.ieee.org
… gradientbased learning technique. Given an appropriate network architecture, gradient-based learning … This paper reviews various methods applied to handwritten character recognition …
☆ 保存 99 引用 被引用次数：63071 相关文章 所有 40 个版本 Web of Science: 28544 »

Very deep convolutional networks for large-scale image recognition
K Simonyan, A Zisserman - arXiv preprint arXiv:1409.1556, 2014 - arxiv.org
… In this work we evaluated very deep convolutional networks (up to 19 weight layers) for largescale image classification. It was demonstrated that the representation depth is beneficial …
☆ 保存 99 引用 被引用次数：119377 相关文章 所有 43 个版本 »

Squeeze-and-excitation networks
J Hu, L Shen, G Sun - … of the IEEE conference on computer …, 2018 - openaccess.thecvf.com
… The role of Excitation. While SE blocks have been empirically shown to improve network performance, we would also like to understand how the self-gating excitation mechanism …
☆ 保存 99 引用 被引用次数：29055 相关文章 所有 30 个版本 Web of Science: 10784 »

Human Error = 5.1%

Reference: https://kaiminghe.github.io/eccv18tutorial/eccv2018_tutorial_kaiminghe.pdf

20

20

Visual Geometry Group (VGG, 2014)

- Smaller kernels and deeper networks.
- 8 layers → 19 layers
- 3*3 CONV stride 1, padding 1, 2*2 MAX pooling stride 2
- 11.7% top-5 error (ZFNet) → 7.3% top-5 error (VGG)

AlexNet

VGG16

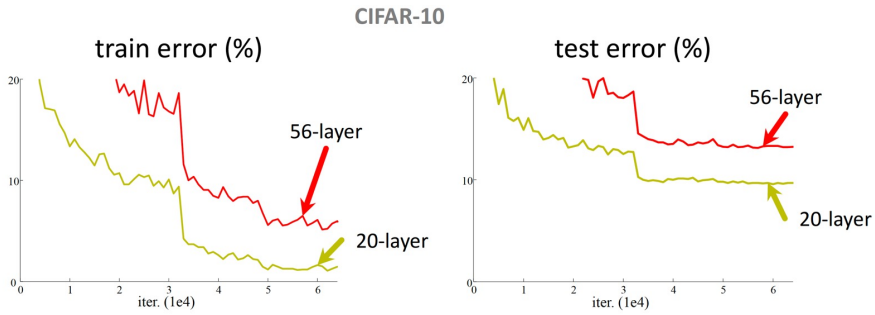
VGG19

Very deep convolutional networks for large-scale image recognition
K Simonyan, A Zisserman - arXiv preprint arXiv:1409.1556, 2014 - arxiv.org
… In this work we evaluated very deep convolutional networks (up to 19 weight layers) for largescale image classification. It was demonstrated that the representation depth is beneficial …
☆ 保存 99 引用 被引用次数：119377 相关文章 所有 43 个版本 »

Reference: http://cs231n.stanford.edu/slides/2023/lecture_6.pdf

ResNet (2016)

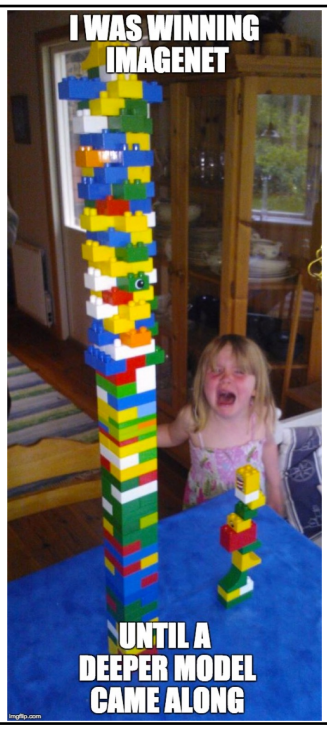
- Why not just stacking the convolution layers?



- Due to vanishing gradient, the training error of 56-layer network is larger than that of 20-layer.
  - Structural estimation with hidden classes.

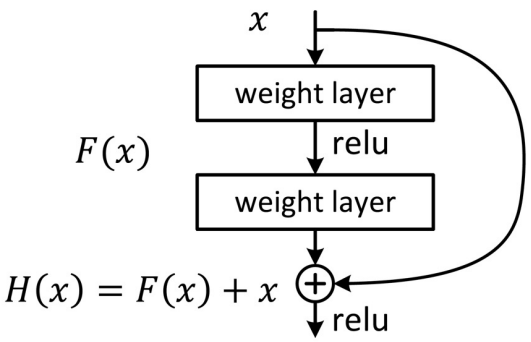
Deep residual learning for image recognition
K He, X Zhang, S Ren, J Sun - Proceedings of the IEEE …, 2016 - openaccess.thecvf.com
… Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. …
★ 保存 99 引用 被引用次数：209436 相关文章 所有 73 个版本 »

Reference: https://kaiminghe.github.io/cvpr18tutorial/cvpr2018_tutorial_kaiminghe.pdf



ResNet (2016)

- Residual net



$H(x)$ is any desired mapping,

~~hope the small subnet fit $H(x)$~~

hope the small subnet fit $F(x)$

let $H(x) = F(x) + x$

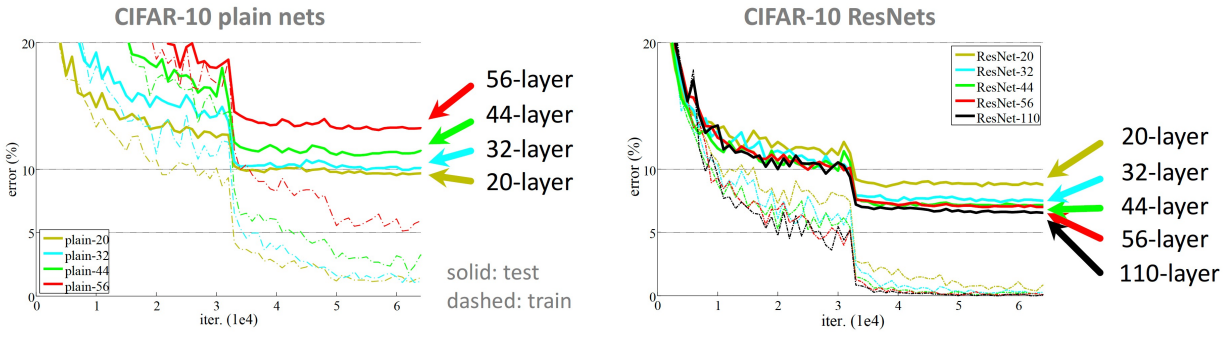
- $F(x)$ is a residual mapping w.r.t. identity

Reference: https://kaiminghe.github.io/cvpr18tutorial/cvpr2018_tutorial_kaiminghe.pdf

23

23

CIFAR-10 Experiments



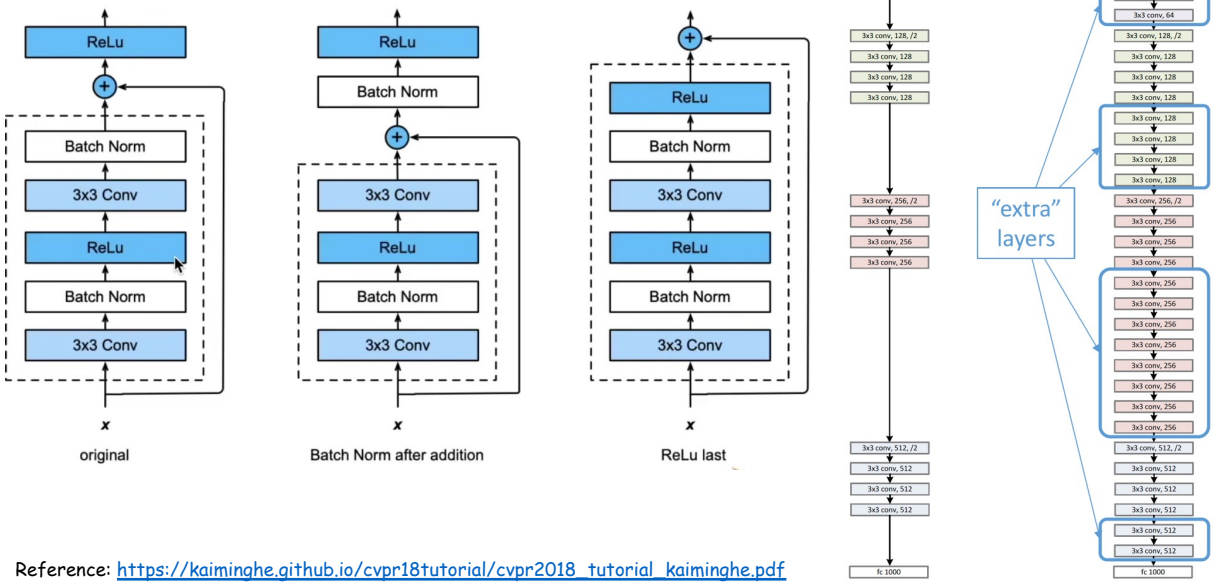
- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

Reference: https://kaiminghe.github.io/cvpr18tutorial/cvpr2018_tutorial_kaiminghe.pdf

24

24

ResNet Blocks and Architecture

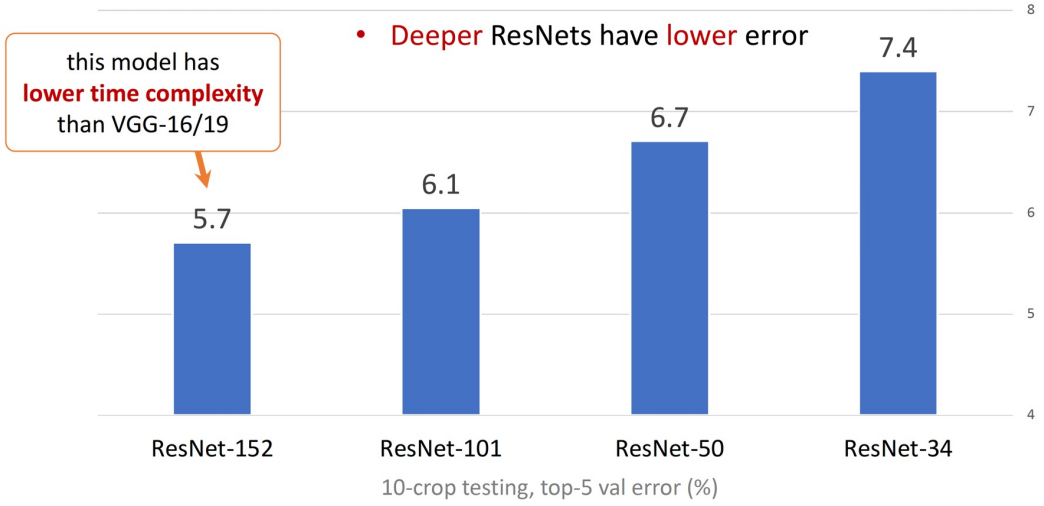


Reference: https://kaiminghe.github.io/cvpr18tutorial/cvpr2018_tutorial_kaiminghe.pdf

25

25

ResNet Performance on ImageNet



ResNet goes much beyond computer vision (recall the transformer architecture)!

Reference: https://kaiminghe.github.io/cvpr18tutorial/cvpr2018_tutorial_kaiminghe.pdf

26

26

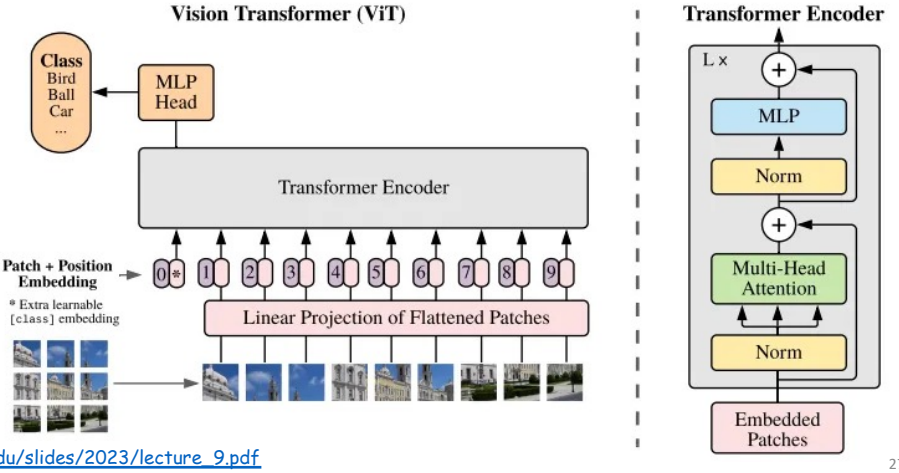
ViT (2020)

An image is worth 16x16 words: Transformers for image recognition at scale
A Dosovitskiy, L Beyer, A Kolesnikov… - arXiv preprint arXiv …, 2020 - arxiv.org
… To do so, we split an image into patches and provide the sequence of … Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image …
☆ 保存 99 引用 被引用次数：32273 相关文章 所有 17 个版本 »

- Let's forget about CNN and use transformer (again☺) encoders to process image.
- Decompose images into patches of 16 by 16 pixels. Patch = token.

Transformer lacks the inductive biases of CNN: Translation invariance and locality.

It may be smart to combine CNN and transformers.



Reference: http://cs231n.stanford.edu/slides/2023/lecture_9.pdf

27

27

ViT vs. ResNet

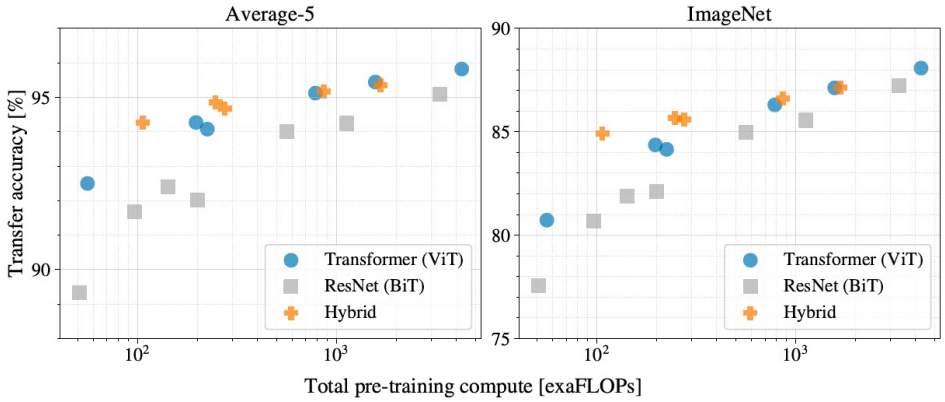


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

Attention is all you need, again!

Reference: http://cs231n.stanford.edu/slides/2023/lecture_9.pdf

28

28

Agenda

- Image Classification
- Image Classification Applications in Business/Econ Research

29

29

ML-Driven Hypothesis Generation

Machine learning as a tool for hypothesis generation
J Ludwig, S Mullainathan
2023 · nber.org

Abstract
While hypothesis testing is a highly formalized activity, hypothesis generation remains largely informal. We propose a systematic procedure to generate novel hypotheses about human behavior, which uses the capacity of machine learning algorithms to notice patterns people might not. We illustrate the procedure with a concrete application: judge decisions about who to jail. We begin with a striking fact: The defendant's face alone matters greatly for the judge's jailing decision. In fact, an algorithm given only the pixels in the defendant's mugshot accounts for up to half of the predictable variation. We develop a procedure that allows human subjects to interact with this black-box algorithm to produce hypotheses about what in the face influences judge decisions. The procedure generates hypotheses that are both interpretable and novel: They are not explained by demographics (eg race) or existing psychology research; nor are they already known (even if tacitly) to people or even experts. Though these results are specific, our procedure is general. It provides a way to produce novel, interpretable hypotheses from any highdimensional dataset (eg cell phones, satellites, online behavior, news headlines, corporate filings, and high-frequency time series). A central tenet of our paper is that hypothesis generation is in and of itself a valuable activity, and hope this encourages future work in this largely "prescientific" stage of science.

nber.org

SHOW LESS

☆ Save 99 Cite Cited by 17 Related articles All 11 versions »

- Directly use CNN algorithms to generate interpretable and testable hypotheses on human behaviors.
- Mug shots + ML predicts judge behavior in jailing decisions, uncovering (about 22.3%) new information never discovered before.
- Create counterfactual mug shots based on the algorithmic discovery and iterate with crowd-sourced workers to confirm what the human judges see is the same as what the ML algorithm sees, thus formalizing the algorithmic discoveries as testable hypotheses.
- There's no causality claimed for these hypotheses.

30

30

What we teach about race and gender: Representation in images and text of children's books

A Adukia, A Eble, E Harrison, HB Runesha, T Szasz

The Quarterly Journal of Economics, 2023 · academic.oup.com

Abstract

Books shape how children learn about society and norms, in part through representation of different characters. We use computational tools to characterize representation in children's books widely read in homes, classrooms, and libraries over the past century and describe economic forces that may contribute to these patterns. We introduce new artificial intelligence methods for systematically converting images into data. We apply these tools, alongside text analysis methods, to measure skin color, race, gender, and age in the content of these books, documenting what has changed and what has endured over time. We find underrepresentation of Black and Latinx people in the most influential books, relative to their population shares, though representation of Black individuals increases over time. Females are also increasingly present but appear less often in text than in images, suggesting greater symbolic inclusion in pictures than substantive inclusion in stories. Characters in these influential books have lighter average skin color than in other books, even after conditioning on race, and children are depicted with lighter skin color than adults on average. We present empirical analysis of related economic behavior to better understand the representation we find in these books. On the demand side, we show that people consume books that center their own identities and that the types of children's books purchased correlate with local political beliefs. On the supply side, we document higher prices for books that center nondominant social identities and fewer copies of these books in libraries that serve predominantly White communities.

Oxford University Press

SHOW LESS

Save Cite Cited by 47 Related articles All 20 versions

Children Book Images

- Data: Award winning children's books from 1920s.
- Use CNN and transfer learning to classify the race, gender and age of the images in children's books.
- Text analysis: Word counts and NER.
- Underrepresentation of Black and Latino in influential Children's books, though the representation of Black increases overtime.
- Empirical analyses investigate the economic behaviors underlying the representations discovered by ML in Children's books.

31

31

Price Trends Prediction from Charts

(Re-) Imag (in) ing Price Trends

J Jiang, B Kelly, D Xiu

The Journal of Finance, 2023 · Wiley Online Library

Abstract

We reconsider trend-based predictability by employing flexible learning methods to identify price patterns that are highly predictive of returns, as opposed to testing predefined patterns like momentum or reversal. Our predictor data are stock-level price charts, allowing us to extract the most predictive price patterns using machine learning image analysis techniques. These patterns differ significantly from commonly analyzed trend signals, yield more accurate return predictions, enable more profitable investment strategies, and demonstrate robustness across specifications. Remarkably, they exhibit context independence, as short-term patterns perform well on longer time scales, and patterns learned from U.S. stocks prove effective in international markets.

Wiley Online Library

SHOW LESS

Save Cite Cited by 25 Related articles All 2 versions

Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation

AW Lo, H Mamaysky, J Wang - The journal of finance, 2000 - Wiley Online Library

Technical analysis, also known as "charting," has been a part of financial practice for many decades, but this discipline has not received the same level of academic scrutiny and acceptance as more traditional approaches such as fundamental analysis. One of the main obstacles is the highly subjective nature of technical analysis—the presence of geometric shapes in historical price charts is often in the eyes of the beholder. In this paper, we propose a systematic and automatic approach to technical pattern recognition using …

Save Cite Cited by 1824 Related articles All 25 versions Web of Science: 523

- Data: Stock-price charts.
- Use CNN to directly extract new price patterns from stock-price charts and yield more accurate return predictions and more profitable investment strategies.

32

32

Poverty Detection with Satellite Images

Combining satellite imagery and machine learning to predict poverty

N Jean, M Burke, M Xie, WM Davis, DB Lobell, S Ermon

Science, 2016 • science.org

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.

AAAS

SHOW LESS

☆ Save 99 Cite Cited by 1663 Related articles All 9 versions Web of Science: 719 »

- Satellite image data + ML to estimate consumption expenditure and asset wealth in Africa.
- Transfer learning: CNN pretrained on ImageNet and finetuned on small-scale labeled data.
- The CNN trained to represent the satellite images explains 75% of economic outcome variations.

33

33

Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp

M Zhang, L Luo

Management Science, 2023 • pubsonline.informs.org

Despite the substantial economic impact of the restaurant industry, large-scale empirical research on restaurant survival has been sparse. We investigate whether consumer-posted photos can serve as a leading indicator of restaurant survival above and beyond reviews, firm characteristics, competitive landscape, and macroconditions. We employ machine learning techniques to extract features from 755,758 photos and 1,121,069 reviews posted on Yelp between 2004 and 2015 for 17,719 U.S. restaurants. We also collect data on restaurant characteristics (e.g., cuisine type, price level) and competitive landscape as well as entry and exit (if applicable) time from each restaurant's Yelp/Facebook page, own website, or Google search engine. Using a predictive XGBoost algorithm, we find that consumer-posted photos are strong predictors of restaurant survival. Interestingly, the informativeness of photos (e.g., the proportion of food photos) relates more to restaurant survival than do photographic attributes (e.g., composition, brightness). Additionally, photos carry more predictive power for independent, young or mid-aged, and medium-priced restaurants. Assuming that restaurant owners possess no knowledge about future photos and reviews, photos can predict restaurant survival for up to three years, whereas reviews are only predictive for one year. We further employ causal forests to facilitate the interpretation of our predictive results. Among photo content variables, the proportion of food photos has the largest positive association with restaurant survival, followed by proportions of outside and interior photos. Among others, the proportion of photos with helpful votes also positively relates to restaurant survival.

This paper was accepted by Juanjuan Zhang, marketing.

Funding: The authors thankNvidia and Clarifai for supporting this research.

Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2022.4359.

INFORMS

SHOW LESS

☆ Save 99 Cite Cited by 72 Related articles All 6 versions Web of Science: 11 »

Restaurant Survival Prediction

- Use CNN to extract 18 features from consumer-generated photos on Yelp.
- Use XGBT to predict restaurant survival.
- Photos are more informative than reviews to predict restaurant survival.
- There's NO way we can interpret the results causally, although causal forests are employed.

34

34

AirBnb Image Quality

What makes a good image? Airbnb demand analytics leveraging interpretable image features

S Zhang, D Lee, PV Singh, K Srinivasan

Management Science, 2022 · pubsonline.informs.org

We study how Airbnb property demand changed after the acquisition of *verified* images (taken by Airbnb's photographers) and explore what makes a good image for an Airbnb property. Using deep learning and difference-in-difference analyses on an Airbnb panel data set spanning 7,423 properties over 16 months, we find that properties with verified images had 8.98% higher occupancy than properties without verified images (images taken by the host). To explore what constitutes a good image for an Airbnb property, we quantify 12 human-interpretable image attributes that pertain to three artistic aspects—composition, color, and the figure-ground relationship—and we find systematic differences between the verified and unverified images. We also predict the relationship between each of the 12 attributes and property demand, and we find that most of the correlations are significant and in the theorized direction. Our results provide actionable insights for both Airbnb photographers and amateur host photographers who wish to optimize their images. Our findings contribute to and bridge the literature on photography and marketing (e.g., staging), which often either ignores the demand side (photography) or does not systematically characterize the images (marketing).

This paper was accepted by Juanjuan Zhang, marketing.

INFORMS

SHOW LESS

Save Cite Cited by 106 Related articles All 5 versions

- Properties with verified images on AirBnb have 8.98% higher occupancy.
- Use CNN to quantify the quality of images on AirBnb in 12 human-interpretable image attributes.
- Verified photos are indeed better w.r.t. these image attributes.

35

35

AirBnb Photo Layout

Estimating and exploiting the impact of photo layout: A structural approach

H Li, D Simchi-Levi, MX Wu, W Zhu

Management Science, 2023 · pubsonline.informs.org

Host-generated property images as a visual channel reveal substantial information about properties. Selecting proper images to display can lead to higher demand and increased rental revenue. In this paper, we define, estimate, and optimize the impacts of Airbnb photos on customers' renting decisions. We apply ResNet-50, a convolutional neural network model, to build two separate, supervised learning models to evaluate the image quality and room types posted by Airbnb hosts. Then, we characterize the overall impacts of photo layout by the room type featured in the photo, photo quality, and order of display on the listings' web pages. To address two estimation challenges in the Airbnb setting, namely, censored demand and changing consideration sets, we propose a novel pairwise comparison model that utilizes customers' booking sequence data to consistently estimate the impact of photo layout on customers' renting decisions. Our estimation results suggest that the cover image has a significantly larger impact than noncover photos and a high-quality bedroom cover image leads to the largest increase in demand. Furthermore, we build a nonlinear integer programming optimization problem and develop an algorithm to determine the optimal photo layout. Our counterfactual analysis suggests that a listing's unilateral adoption of optimal photo layout leads to 11.0% more bookings on average. Moreover, depending on the neighborhood and market size, when listings simultaneously switch to the optimal photo layout, they get booked for two to five additional days in a year on average, which boosts revenue by $500 to $1,100.

This paper was accepted by Swaminathan, Jayashankar, operations management.

Funding: This research was partially sponsored by the MIT Data Science Lab and also benefited from generous support provided by Zalando.

Supplemental Material: The online companion and data are available at https://doi.org/10.1287/mnsc.2022.4616.

INFORMS

SHOW LESS

Save Cite Cited by 10 Related articles All 7 versions Web of Science: 2

- Use ResNet-50 to predict the human-labeled scores of the AirBnb photos.
- Use a "structural model" to describe how photo quality scores impact consumer rent behaviors.
- Study the "optimal" counterfactual photo layout strategy.

36

36