How to Protect Intellectual Property of Deep Neural Networks

Nicolas Robertson
Computer Science
Senior Seminar
University of Minnesota Morris
April 2021

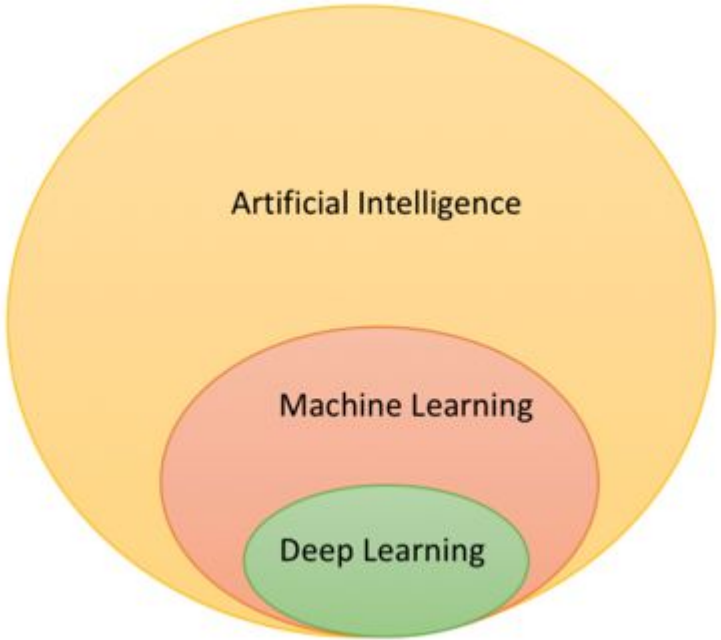
Introduction

- Alice creates a deep neural network
  - Intends to sell to Bob
- Oscar, Bob and Eve infringe on Alice's rights
  - Oscar wants to make a copy of Alice's model
  - Bob resells model to Eve
  - Eve steals model from Bob
- How can Alice protect herself?

Outline

1. What are deep neural networks?
2. Threat to security and legality
3. What are digital watermarks?
4. Generative Adversarial Networks
5. IPP Blind-Watermark Framework (Li et al)
6. Results
    a. Evaluation of IPP Framework under 5 criteria

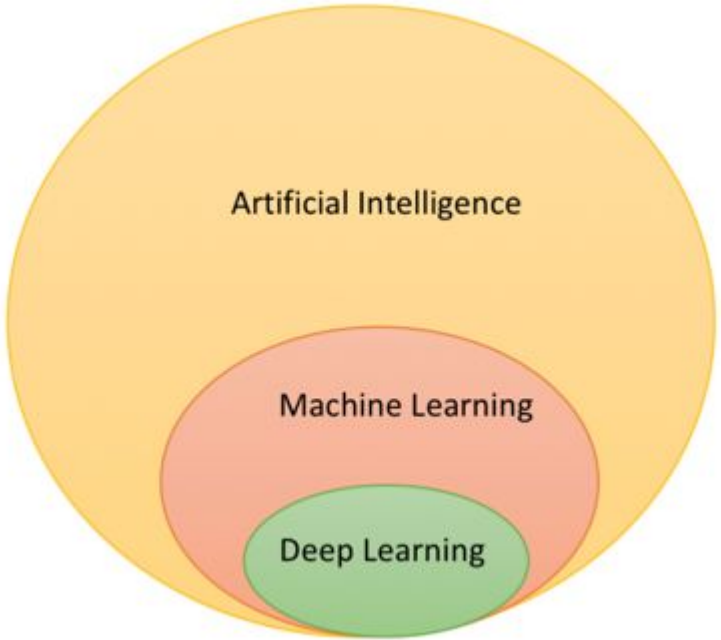
Artificial Intelligence vs Machine Learning



https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb

- Artificial Intelligence (AI)
  - Computers making decisions
  - Explicitly programmed
- Machine Learning
  - Computers learning to make decisions
  - NOT explicitly programmed

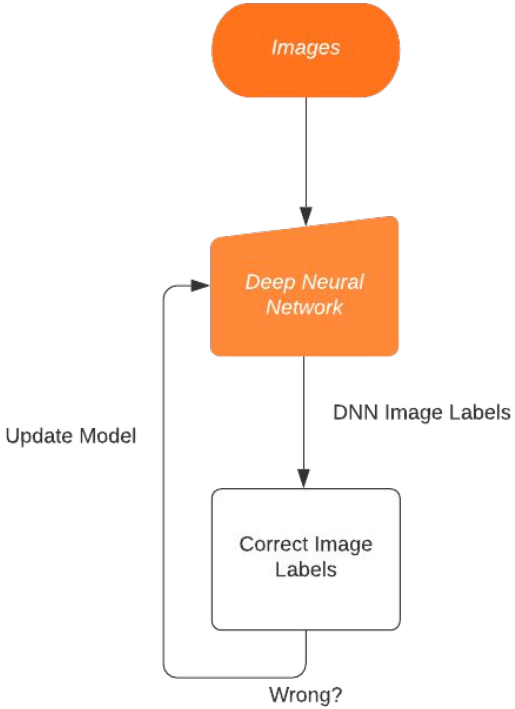
What are deep neural networks (DNN)?



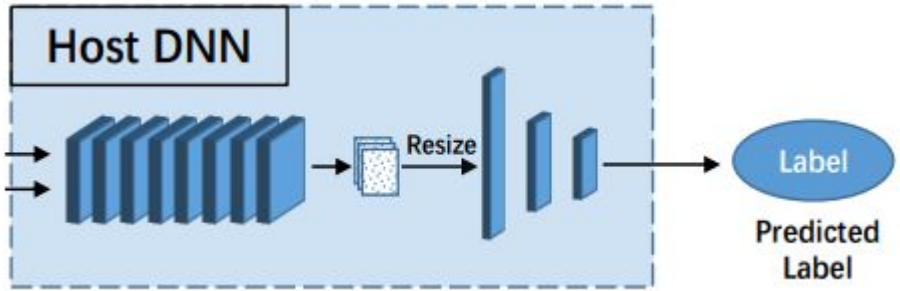
https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb

- Traditional machine learning struggles with certain tasks
- Deep Learning
  - Can handle such tasks
- Requires deep neural networks (DNN)
  - Mimics human brain

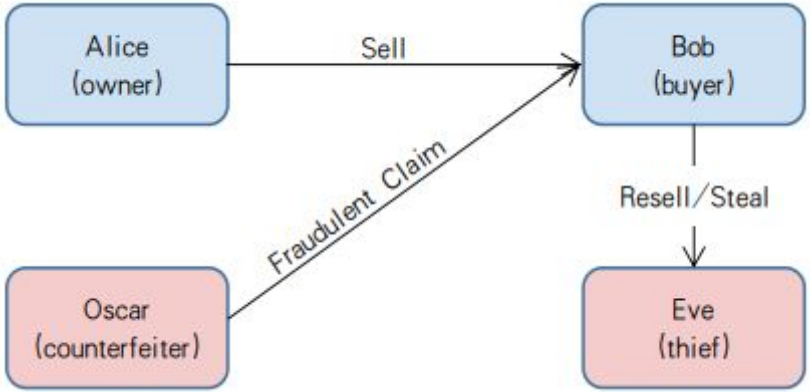
How do deep neural networks learn (Training)?



- DNN classifies data
- Classification happens in layers
- Adjust DNN after each process
- Goal: Achieve perfect classification
- Layers
  - More = more computational power



Threat to security and legality



https://dl.acm.org/doi/10.1145/3359789.3359801

- Evasion Attack
  - Eve steals DNN
  - Bob resells DNN without permission
- Fraudulent Claim of Ownership

Digital Watermarks

- Confirm authenticity and ownership of intellectual property
- Involves embedding data

Ordinary Sample (Original Image)

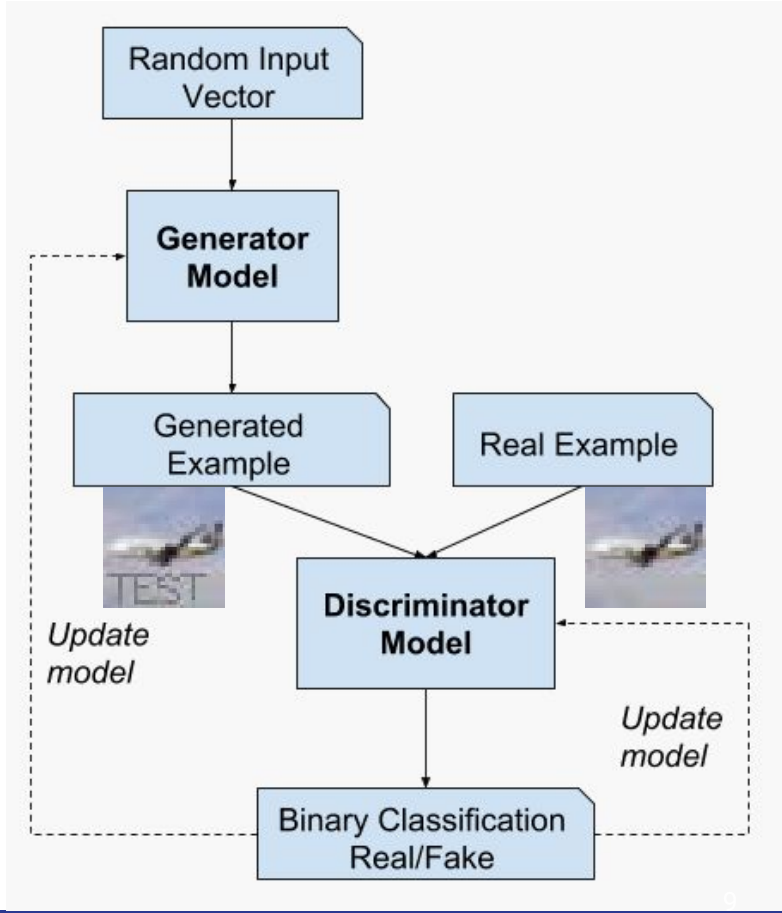
Key Sample (Watermarked Image)

https://dl.acm.org/doi/10.1145/3359789.3359801

Generative Adversarial Networks (GANS)

Model = Neural Network

- Two player game
- Step 1: Generator
  - Generates key samples
- Step 2: Discriminator
  - Identifies key samples
- Step 3: Update Models
- Goal: **Fool Discriminator**

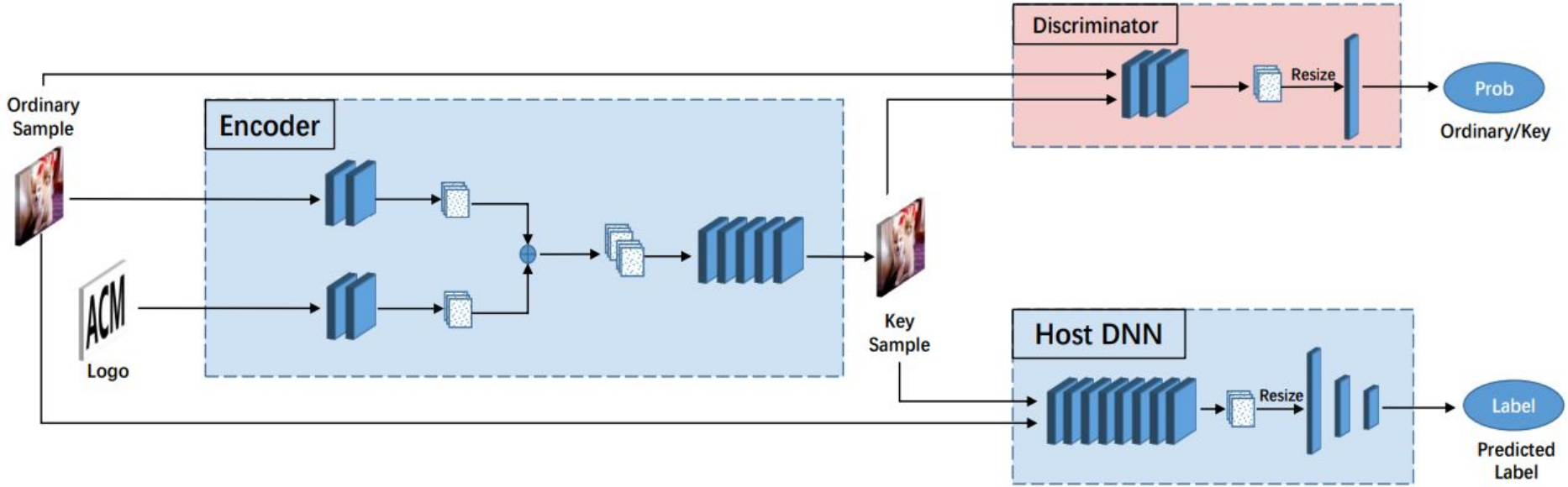


https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

IPP Blind-Watermark Framework

- Authors: Li et al
- Uses GANs to strengthen IP protection
  - Make key sample watermarks near-invisible

Li et al's IPP Framework



https://dl.acm.org/doi/10.1145/3359789.3359801

IPP Blind-Watermark Framework

- IPP watermarks a DNN
- DNN will recognize key samples
  - Proves Alice’s ownership
  - This behavior serves as the watermark

Results

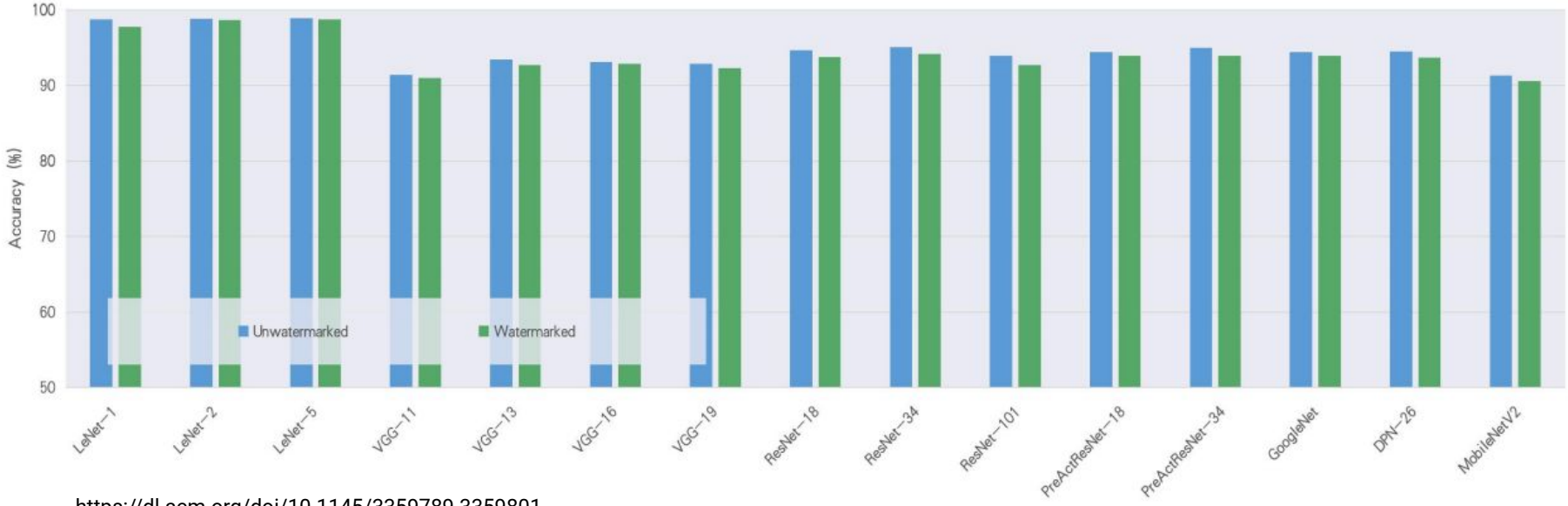
Evaluation on 5 criteria

- Fidelity
- Effectiveness & Integrity
- Security
- Legality
- Feasibility

Fidelity

Is the primary classification task affected by side effects?

Fidelity

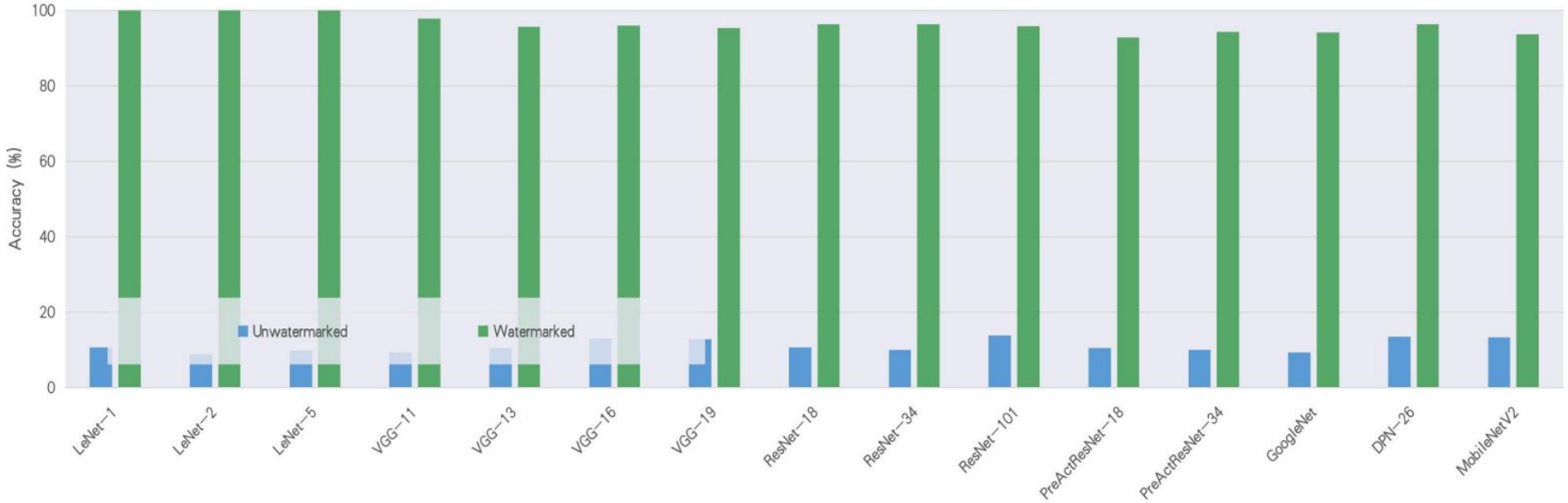


https://dl.acm.org/doi/10.1145/3359789.3359801

Effectiveness & Integrity

How successfully can the watermark verify the host DNN?

Effectiveness and Integrity



https://dl.acm.org/doi/10.1145/3359789.3359801

Security

How well can it defend against evasion attack?

Security

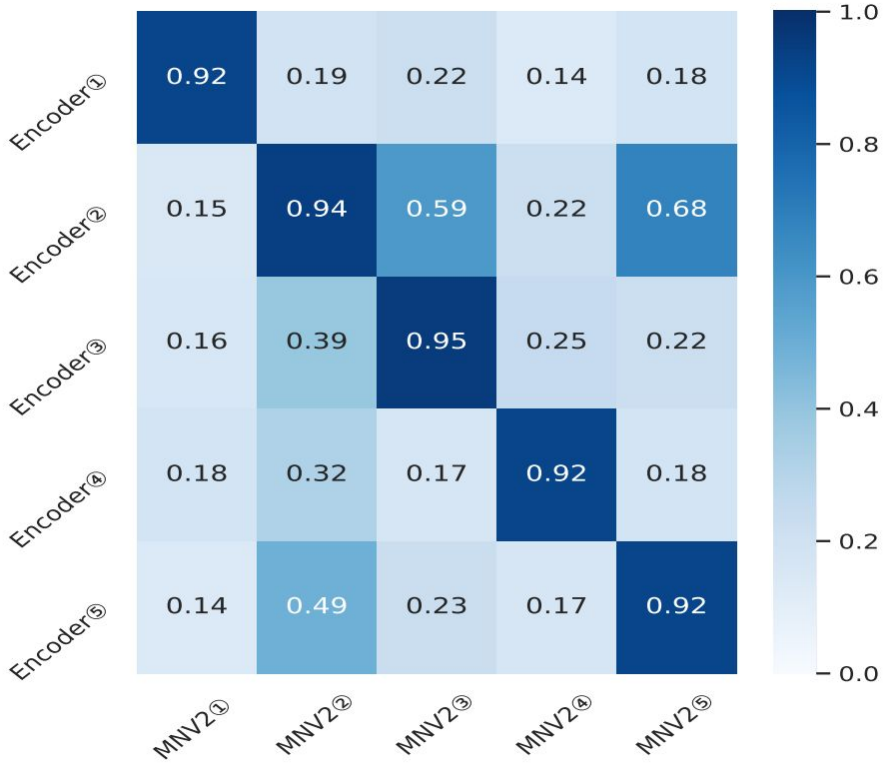
https://dl.acm.org/doi/10.1145/3359789.3359801

Legality

Can it defend against counterfeiting?

Legality

- What if the key samples leaked?
  - Oscar makes his own key samples
- Test against counterfeiting?
  - Train 5 instances of IPP with different key samples
  - Test 5 instances of MNV2 watermarked with different IPP instances



https://dl.acm.org/doi/10.1145/3359789.3359801

Feasibility

Does it clearly associate a DNN with its author?

Feasibility

- Overfitting
  - Classification too used to training set
    - Doesn’t work on other sets
  - IPP overfitted on key samples
  - Ensures association between DNN and author

Conclusion

Fidelity	√
Effectiveness	√
Integrity	√
Security	√
Legality	√
Feasibility	√

- IPP framework meets all requirements
- Other watermark techniques fail one or more requirements

https://dl.acm.org/doi/10.1145/3359789.3359801

Questions?

References

- *How to Prove Your Model Belongs to You: A Blind-Watermark Based Framework to Protect Intellectual Property of DNN*. Li, Zheng and Hu, Chengyu and Zhang, Yang and Guo, Shanqing. Proceedings of the 35th Annual Computer Security Applications Conference, 2019. https://dl.acm.org/doi/10.1145/3359789.3359801
- *A Layman's Guide to Deep Neural Networks*. Moolayil. 2019. https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb
- *A Gentle Introduction to Generative Adversarial Networks (GANs)*. Jason Brownlee. 2019. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/