First Quantization Matrix Estimation for DJPEG Images: An Adversarial Analysis

Manish Okade$^{[0000-0003-1500-2693]}$

National Institute of Technology, Rourkela, Odisha, India
okadem@nitrkl.ac.in

Abstract. This paper investigates the role of an intelligent adversary in derailing the advantages offered by machine and deep learning algorithms. The first quantization matrix estimation (FQME) forensic research problem in Double JPEG (DJPEG) compressed images is chosen as an example case study to demonstrate the vulnerabilities of the machine and deep learning algorithms that an intelligent adversary can exploit. DJPEG compression involves two compression cycles: the first compression when the image is initially saved as JPEG and the second compression when a forger manipulates the image and again saves it in the JPEG format. In such cases, the information regarding the first compression is lost in the presence of the second JPEG compression. Specifically, the quantization coefficients are of interest since estimating the quantization coefficients for the first compression, often referred to as the primary quantization estimation, can give information about the history of the image and the possibility of forgery/tampering. Various methods exist for estimating the first quantization coefficients, both statistical and deep learning-based. However, existing works do not evaluate the robustness of these estimation models against adversarial attacks, which is an essential criterion from a security point of view. In this work, a comprehensive adversarial analysis is carried out to show the vulnerabilities of machine and deep learning models for the FQME forensic research problem. Such a detailed security analysis is the need of the hour, and this paper throws light on it from a forensic perspective.

Keywords: Adversarial Attack · DJPEG · Image Forensics · FQME · Adversarial defence.

1 Introduction

Multimedia data in today's world are ubiquitous. The internet is full of audio clips, images and videos that are widely shared on many social media platforms. In many instances, there are integrity concerns regarding the multimedia data being circulated. Multimedia forensics focuses on answering integrity concerns by detecting such manipulations. Of particular interest in this paper is image forensics, especially JPEG forensics, which focuses on detecting the clues left over by the forger in a JPEG scenario since it is a widely used image compression standard. In JPEG forensics, an interesting case is that of Double JPEG

(DJPEG), where a JPEG compressed image is decompressed to replace a certain image region with content from another image, followed by resaving the image as a JPEG image. As a result of two JPEG compressions, there would be a presence of two quantization matrices: the first quantization matrix when the compression took place initially and the second quantization matrix that would be present after the image is possibly forged and resaved. The second compression cycle erases the first quantization matrix. Hence, for evaluating the authenticity, the first quantization matrix needs to be estimated, referred to as the first quantization matrix estimation (FQME) forensic research problem in Double JPEG (DJPEG) compressed images. Fig. 1 shows the complete JPEG pipeline. To apply DCT transformation, the image is partitioned into 8×8 blocks. The below equation gives the DCT transform that is applied blockwise.

$$DCT(u,v) = \frac{1}{\sqrt{2N}} C(u)C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) cos \left[\frac{(2x+1)u\pi}{2N} \right] cos \left[\frac{(2y+1)v\pi}{2N} \right] \tag{1}$$

where $N = 8$ and $C(x) = \frac{1}{\sqrt{2}}$ if $x = 0$ else $1 if x > 0$. In the quantization step, the transformed DCT coefficients are quantized using pre-defined quantization tables specified by the JPEG standard. The 8×8 matrix is denoted by 'Q', which stands for the quantization matrix used in the quantization step of the JPEG standard. When a raw image is subjected to a JPEG compression using quantization matrix Q_1, decompressed with the possible intent of modification, and then compressed back to a JPEG format using quantization matrix Q_2, a double compressed JPEG image is obtained. Q_1 and Q_2 contain 64 coefficients each. The JPEG quality factor is referred to as 'QF'. A 'QF' value specifies a quantization matrix Q according to the JPEG standard. The second compression quality factor is denoted by QF_2, and the first with QF_1.

Researchers have investigated several techniques for the FQME forensic research problem in DJPEG images both conventional and, more recently, deep learning-based. With an increasing focus on deep learning-based methods due to high model accuracy and being end-to-end, they have brought out a lot of advantages that conventional techniques lack, making them the preferred choice of algorithm development. However, a major security threat exists for deep learning models, i.e. the possible presence of an intelligent adversary, which, if it remains unaccounted, could pose severe security threats. Adversaries can easily fool the deep learning models by adding perturbations to the input using various methods. The perturbations, nearly invisible to the human eye, can cause a major performance drop for deep learning models. The paper is a novel attempt to study the adversarial attacks in detail when applied to the FQME forensic research problem in DJPEG images. Alongside adversarial defences are also proposed to mitigate the ill effects of an adversarial attack launched by an informed adversary.

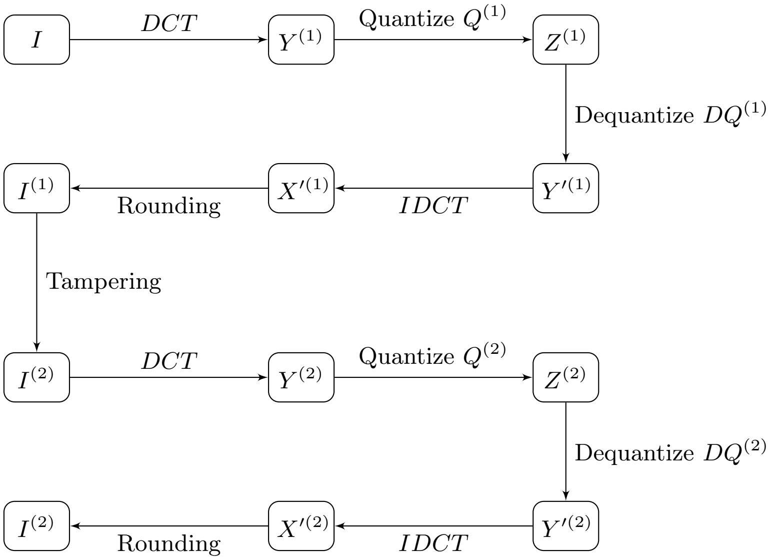

Fig. 1. Block diagram of Double JPEG compression process with tampering operation between the $first^{(1)}$ JPEG and *second*$^{(2)}$ JPEG compression cycles.

2 Literature Review

Conventional works on FQME [6, 7, 4, 3] are either DCT domain-based or statistics-based. Farid [6] proposed a forgery detection method by performing a third quantization step and computing the error between DCT coefficients, thereby locating two minima corresponding to the first and second quantization. Galvan et al. [7] improved this method by modelling the split and residual noise and mitigating them via a DCT histogram-based filtering approach, which resulted in accurate first quantization matrix coefficients. However, their DCT histogram filtering approach failed for special cases of split noise. This was improved in [4], where a novel priority assignment and selection strategy was applied to error function values of the second quantisation step's existing and missing multiples, which achieved accurate FQME. Recently, deep learning-based models have gained popularity in image forensics due to the higher accuracy that they offer and their ability to be trained in an end-to-end fashion.

Niu et al. [11] proposed a deep learning model by modifying the dense-net architecture for the FQME forensic research problem. The method achieved good accuracy for all quality factor pairs, and good generalization performance was demonstrated by testing for quality factors different from those in the training stage and on images from different datasets. However, no adversarial analysis [9, 14, 13] was carried out, leaving the model vulnerable to attacks by an informed adversary. Along similar lines, Battiato et al. [2] also carried out FQME by utilizing a CNN-based model that employed 1-D histograms of DCT values coupled with a regularization term to improve the estimation accuracy. Still, this model, too, lacked adversarial analysis. A survey of existing problems that have received adversarial attention is available in [12]. It is evident from the survey that although many research problems, like face recognition [1], gait recognition [10], etc., have received attention from an adversarial point of view, the FQME forensic research problem is still left untouched in terms of comprehensive adversarial analysis. Therefore, existing deep learning models for FQME have not been tested under adversarial settings, leaving enough scope for research in the interesting domain of adversarial attacks and defences. Moreover, in image forensics, along with achieving good accuracy, it is also crucial to consider the deep learning model's robustness. While Niu et al. [11] model addressed the model's generalization ability across datasets and quality factors, the model's performance was not studied from the security point of view. Hence, it becomes paramount to understand if the model would continue to show good performance under an adversarial attack. This paper is motivated by the premise of studying the performance of the existing deep learning-based FQME models in an adversarial setting, along with proposing defences to mitigate the possible attacks, thereby achieving secure FQME models.

3 Key Contributions

The key contributions are listed below.

- Detailed performance analysis of existing deep learning-based models for the FQME forensic research problem under an adversarial setting. Two types of adversarial attacks, the Limited-Broyden Fletcher Goldfarb Shanno (L-BFGS) attack and the Fast Gradient Sign Method (FGSM) attacks are launched on existing deep learning-based FQME models to analyze the model's robustness. Inferences are drawn that help in designing good defences.
- Adversarial defences like adversarial retraining are explored to mitigate the drawbacks of the launched adversarial attacks on existing deep learning-based models for the FQME forensic research problem. Overall, such a detailed security analysis of the FQME forensic research problem forms a novel contribution which was lacking in the literature.
- The marriage of forensics with security outlined in the paper, which brings trust in the credibility of image data, is the current need of the hour in the domain of information forensics and security. The paper throws light on these aspects.

4 Adversarial attacks and defences: FQME in DJPEG images

The section is devoted towards a comprehensive security analysis of the FQME forensic research problem in DJPEG images. The first subsection is devoted to understanding the performance of existing deep learning-based methods when adversarially attacked, while the next subsection explores mitigation measures in the form of adversarial defences, also referred to as adversarial countermeasures, that can overcome the attack drawbacks.

4.1 Adversarial attacks

As an example test case, two existing deep learning-based FQME models are considered, i.e. Niu et al. [11] model, and Battiato et al. [2] model. The two FQME models are adversarially attacked using the Limited-Broyden Fletcher Goldfarb Shanno (L-BFGS) and Fast Gradient Sign Method (FGSM) attacks. The two attacks are chosen to demonstrate the performance drop in the existing deep learning-based FQME models.

L-BFGS Attack: It was the first among the attacks introduced against deep neural networks. L-BFGS attack [13, 9, 14] is an optimization algorithm belonging to the broad family of quasi-Newton methods. It works by generating adversarial examples using an L-BFGS method to solve the general targeted problem given below, which involves calculating approximate values of adversarial examples by line-searching c > 0.

$$\begin{aligned} \min_{r} \quad & c\|r\|_p + J_\theta(x+r,t) \\ \text{st} \quad & x+r \in [0,1] \end{aligned} \tag{2}$$

where 'r' represents the perturbation vector, 'x' denotes the original input, and 't' is the target label. J_θ is the loss function, e.g., the cross entropy, 'c' is a suitable constant which finds a compromise between the perturbation magnitude and the attack performance. The L-BFGS attack uses the second-order quasi-Newton method to solve this problem.

FGSM attack: The Fast Gradient Sign Method (FGSM) proposed by Goodfellow et al. [9] is a fast method to generate adversarial examples since the L-BFGS attack was slow due to the utilization of an expensive linear search method to find the optimal value. FGSM performs one step gradient update along the direction of the gradient sign at each pixel, as given below.

$$x' = x + \epsilon * sign[\nabla_x J(\theta, x, y)] \tag{3}$$

where 'x' is the image input, $J(\theta, x, y)$ is the cost function, and $\nabla_x J(\theta, x, y)$ is the gradient of the cost function with respect to the input. The value of ϵ controls the magnitude of perturbation to be added to the image.

4.2 Adversarial defence

Adversarial defences or countermeasures are basically of two types: reactive and proactive. The former detects adversarial examples after the deep learning model is built, while the latter makes deep neural networks more robust before adversaries generate adversarial examples. Proactive defences have more importance since they account for the presence of an intelligent adversary well in advance. Adversarial retraining is one of the proactive defence strategies explored in this paper for the FQME deep learning models. As outlined earlier, the L-BFGS and FGSM attacks dent the accuracy scores of the existing state-of-the-art FQME deep learning models since they were trained in a non-adversarial setting. In this section, adversarial retraining is adopted in training Niu et al. [11] model and Battiato et al. [2] model. By generating adversarial examples in every step of training and injecting them into the training set of Niu et al. [11] model and Battiato et al. [2] model, it is observed that the adversarial retraining improves the robustness of deep neural networks. The improvement in accuracy achieved for the two state-of-the-art models is shown in the results section. Additionally, by incorporating adversarial retraining, the precision of the two models is improved as it provides some implicit regularization for deep neural networks.

5 Experimental Results

Images for experimentation are taken from two state-of-the-art datasets utilized in image forensics, namely the RAISE dataset [5] and the DRESDEN dataset [8]. In order to have a fair comparison, the experimental methodology followed by the two existing FQME models, i.e. Niu et al. [11] model and Battiato et al. [2] model, are kept unchanged. In other words, the data preparation methodology of both models is followed in letter and spirit. The first 15 quantization matrix coefficients were estimated using both methods in line with prior works. Niu et al. [11] model utilized two cases for training; $QF_1 \in \{60, 65, 70, 75, 80, 85, 90, 95, 98\}$ for $QF_2 = 90$ and $QF_1 \in \{55, 60, 65, 70, 75, 80, 85, 90, 95\}$ for $QF_2 = 80$. 4,00,000 image patches per QF_1 were used for training, and 7440 image patches were taken for testing. Battiato et al. [2] model also utilized $QF_2 = 90$ and $QF_2 = 80$ as Niu et al. [11] and QF_1 as earlier for uniformity. Battiato et al. [2] achieved superiority in their model by including a regularization term that minimized the differences among neighbouring first quantization values. However, neither method accounts for an informed adversary and, as a result, lacks security aspects, which is of paramount importance and the focus of the work reported in this paper.

5.1 Adversarial attack analysis

Two types of adversarial attacks, i.e. L-BFGS and FGSM, are launched on Niu et al. [11] method and Battiato et al. [2] method to test the robustness of the FQME models. A set of 1000 images is chosen randomly from the RAISE and DRESDEN datasets. Adversarial examples are generated using the L-BFGS attack method by line-searching c > 0, which generates the perturbed test images. While generating the perturbed samples in an L-BFGS attack, it is observed that the method is slow because it is designed to find the smallest possible attack perturbation. On similar lines, for the FGSM attack, ϵ is chosen as '0.005' to be in the range $\epsilon \in [0, 0.01]$. For each of these images, the perturbation value

Table 1. Accuracy scores for 'No Attack', L-BFGS and FGSM attack with QF_2=90.

QF_1	No Attack		L-BFGS attack		FGSM attack	
	Niu's model [11]	Battiato's model [2]	Niu's model[11]	Battiato's model[2]	Niu's model[11]	Battiato's model[2]
55	0.00	0.80	0.00	0.55	0.00	0.42
60	0.64	0.86	0.45	0.52	0.36	0.41
65	0.54	0.84	0.37	0.50	0.28	0.40
70	0.66	0.88	0.44	0.53	0.37	0.42
75	0.77	0.89	0.49	0.55	0.38	0.43
80	0.81	0.87	0.52	0.51	0.41	0.42
85	0.81	0.90	0.53	0.57	0.42	0.45
90	0.02	0.12	0.00	0.00	0.00	0.00
95	0.78	0.62	0.48	0.45	0.36	0.34
98	0.76	0.76	0.47	0.48	0.35	0.37
Mean	0.58	0.75	0.38	0.46	0.29	0.36

is calculated by utilizing the sign of the gradient of the cost function with respect to the input, multiplied by the ϵ value to control the magnitude, i.e. given in Equ. (3). This perturbation is then added to the original image to generate adversarial images.

Niu et al. [11] and Battiato et al. [2] models are tested on the generated adversarial images to obtain the accuracy scores. Table 1 and Table 2 show the accuracy scores obtained for the two adversarial attack types when applied to the existing state-of-the-art FQME models for QF_2=90 and QF_2=80, respectively. As observed, there is a drop in accuracy for both methods in general when adversarial images are utilized to test the built models, signifying vulnerability to attacks from an informed adversary. In particular, it is observed from the two tables that the FGSM attack is stronger than the L-BFGS attack for QF_2=90 and QF_2=80. This is due to parameter ϵ, which controls the magnitude of the perturbation in the FGSM attack. Another important observation from the two tables is that QF_2=90 performs better than QF_2=80 in no attack and the two attack types. This is because lower QF_2 values are detrimental to the FQME process due to heavier post-compression, as observed from the mean accuracy scores.

5.2 Adversarial Defence: Retraining

As observed in the earlier subsection, the performance of existing FQME models deteriorates in the presence of an informed adversary, i.e. L-BFGS and FGSM attacks. The role of adversarial defence or countermeasures is to counter and mitigate the effects of such an informed attack. In this paper, adversarial retraining is the defence that is chosen among the various defences to mitigate adversarial attacks. The motivation for choosing adversarial retraining stems from the fact

Table 2. Accuracy scores for 'No Attack', L-BFGS and FGSM attack with QF_2=80

QF_1	No Attack		L-BFGS attack		FGSM attack	
	Niu's model [11]	Battiato's model [2]	Niu's model[11]	Battiato's model[2]	Niu's model[11]	Battiato's model[2]
55	0.24	0.61	0.11	0.55	0.07	0.42
60	0.50	0.65	0.39	0.57	0.28	0.44
65	0.31	0.71	0.22	0.60	0.13	0.49
70	0.50	0.82	0.39	0.71	0.28	0.59
75	0.15	0.65	0.06	0.54	0.00	0.43
80	0.00	0.05	0.00	0.00	0.00	0.00
85	0.04	0.53	0.00	0.41	0.00	0.30
90	0.48	0.40	0.36	0.29	0.24	0.18
95	0.95	0.35	0.77	0.22	0.61	0.14
98	0.21	0.43	0.09	0.31	0.00	0.19
Mean	0.33	0.52	0.23	0.42	0.16	0.31

that it is intuitive and also has good attack mitigation capabilities, especially for the L-BFGS and FGSM attacks. Adversarial retraining works on the concept of including adversarial samples in the training dataset during model training. This helps the model learn in the presence of adversarial samples. More details of adversarial retraining can be referred to in [13].

Adversarial samples are included in the training set, and both the models Niu et al. [11] and Battiato et al. [2] are retrained. The retraining is carried out by exactly following the training steps outlined earlier, with the only difference being that the training set is no longer pristine but contains the perturbed adversarial images. Table 3 shows the performance analysis after incorporating the adversarial retraining defence strategy. As observed, the accuracy drop is mitigated for L-BFGS and FGSM attacks when adversarial retraining is adopted as a countermeasure, i.e. there is an improvement in the accuracy score values for the two models in both the attack cases. However, it is to be noted that the adversarial retrained models are sub-optimal when compared with the no-attack scenario, signifying that accounting for the presence of an informed adversary does have an overhead. In other words, it is to be noted that the no-attack scenario achieves the best accuracy in comparison to the adversarial defence scenario, which is the price attributed to incorporating adversarial analysis. However, the price paid for incorporating adversarial analysis is still acceptable since the deep learning models are more resilient to attacks, which is the need of the hour.

6 Conclusions

This paper provides a detailed adversarial analysis of existing deep learning-based FQME models. Two types of adversarial attacks, namely L-BFGS and

Table 3. Accuracy scores for adversarial retaining defence strategy

QF_1	QF_2=90 (Adversarial retraining)				QF_2=80 (Adversarial retraining)			
	L-BFGS		FGSM		L-BFGS		FGSM	
	Niu's model [11]	Battiato's model [2]	Niu's model [11]	Battiato's model [2]	Niu's model [11]	Battiato's model [2]	Niu's model [11]	Battiato's model [2]
55	0.00	0.61	0.00	0.66	0.18	0.58	0.22	0.55
60	0.55	0.65	0.53	0.71	0.44	0.61	0.46	0.57
65	0.47	0.71	0.42	0.70	0.27	0.66	0.27	0.66
70	0.55	0.82	0.51	0.74	0.44	0.75	0.45	0.75
75	0.65	0.65	0.62	0.76	0.11	0.61	0.11	0.56
80	0.72	0.05	0.70	0.74	0.00	0.00	0.00	0.00
85	0.77	0.53	0.70	0.79	0.02	0.47	0.02	0.42
90	0.01	0.40	0.01	0.00	0.45	0.37	0.45	0.35
95	0.66	0.35	0.67	0.51	0.86	0.30	0.84	0.29
98	0.65	0.43	0.66	0.65	0.17	0.38	0.17	0.36
Mean	0.50	0.52	0.48	0.65	0.24	0.61	0.29	0.45

FGSM, were studied along with adversarial retraining countermeasures to mitigate the two attack types. It was shown that existing FQME models were vulnerable to attacks from an intelligent adversary, thereby calling researchers in the domain of forensics and security to propose models accounting for the presence of adversaries. Among the two attacks, FGSM was found to be more effective as it was fast along with being deceptive in comparison to L-BFGS. The tunable ϵ parameter in the FGSM attack rendered it more effective than the L-BFGS. Adversarial retraining was explored as a defence strategy to mitigate the two attack types, which provided good resilience. However, as expected, the FQME models could not achieve no-attack accuracy when the retraining defence was applied, signifying the price paid for achieving resilience. Although there was a drop in accuracy for adversary-aware models, as seen in the countermeasure subsection, researchers can trade off the drop in accuracy to achieve more secure models. The paper highlighted the importance of achieving security coupled with good accuracy for deep learning-based FQME models, which had not received enough attention in the literature.

Acknowledgments. This work is supported by ANRF (formerly SERB), Government of India, under grant number MTR/2023/000106.

References

1. Alparslan, Y., Keim-Shenk, J., Khade, S., Greenstadt, R.: Adversarial attacks on convolutional neural networks in facial recognition domain. CoRR **abs/2001.11137** (2020), https://arxiv.org/abs/2001.11137

2. Battiato, S., Giudice, O., Guarnera, F., Puglisi, G.: Cnn-based first quantization estimation of double compressed jpeg images. Journal of Visual Communication and Image Representation **89**, 103635 (2022). https://doi.org/https://doi.org/10.1016/j.jvcir.2022.103635, https://www.sciencedirect.com/science/article/pii/S1047320322001559
3. Bianchi, T., Piva, A.: Image forgery localization via block-grained analysis of jpeg artifacts. IEEE Transactions on Information Forensics and Security **7**, 1003–1017 (06 2012). https://doi.org/10.1109/TIFS.2012.2187516
4. Dalmia, N., Okade, M.: Robust first quantization matrix estimation based on filtering of recompression artifacts for non-aligned double compressed jpeg images. Signal Processing: Image Communication **61** (11 2017). https://doi.org/10.1016/j.image.2017.10.011
5. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G.: RAISE: A raw images dataset for digital image forensics. In: Proceedings of the 6th ACM Multimedia Systems Conference. p. 219–224. MMSys '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2713168.2713194, https://doi.org/10.1145/2713168.2713194
6. Farid, H.: Exposing digital forgeries from jpeg ghosts. IEEE Transactions on Information Forensics and Security **4**(1), 154–160 (2009). https://doi.org/10.1109/TIFS.2008.2012215
7. Galvan, F., Puglisi, G., Bruna, A.R., Battiato, S.: First quantization matrix estimation from double compressed jpeg images. IEEE Transactions on Information Forensics and Security **9**(8), 1299–1310 (2014). https://doi.org/10.1109/TIFS.2014.2330312
8. Gloe, T., Böhme, R.: The dresden image database for benchmarking digital image forensics. Journal of Digital Forensic Practice **3**(2-4), 150–159 (2010). https://doi.org/10.1080/15567281.2010.531500, https://doi.org/10.1080/15567281.2010.531500
9. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv 1412.6572 (12 2014)
10. Jia, M., Yang, H., Huang, D., Wang, Y.: Attacking gait recognition systems via silhouette guided gans (2019), https://api.semanticscholar.org/CorpusID:204837725
11. Niu, Y., Tondi, B., Zhao, Y., Barni, M.: Primary quantization matrix estimation of double compressed jpeg images via cnn. IEEE Signal Processing Letters **27**, 191–195 (2020). https://doi.org/10.1109/LSP.2019.2962997
12. Ozdag, M.: Adversarial attacks and defenses against deep neural networks: A survey. Procedia Computer Science **140**, 152–161 (2018), cyber Physical Systems and Deep Learning Chicago, Illinois November 5-7, 2018
13. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014), http://arxiv.org/abs/1312.6199
14. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning (2018)