UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS

KNOWLEDGE DISTILLATION FOR COMPACT FACIAL LANDMARK DETECTION

CS680 Project Report

Winter 2021

AUTHORS

Ankit Shukla(20828606), Mingzhe Huang(4944090)
Candidates for Master of Data Science and Artificial Intelligence
{a27shukl, m78huang}@uwaterloo.ca

INSTRUCTOR

Gautam Kamath, Ph.D.
Assistant Professor, David R. Cheriton School of Computer Science

April 24, 2021

Contents

1 Introduction

Compact neural-networks with lower memory footprint and computational cost are needed for hardware-constrained mobile hardware. In such a case, reduction in network parameters is a necessity, but it is a known fact that parameter reduction may highly impact the model performance. The proposed work intends to explore various knowledge-distillation(KD) based techniques to design a compact network for a common computer vision module - facial landmark detection. This work would attempt to achieve acceptable better performance for a lightweight network, over a non-KD version of the same network trained on the same dataset.

Facial landmark detection was chosen as the target for the proposed work, as it has widely used in many face-related applications [1–4]. It is essentially estimating facial key points around identifying facial features like eyes, eyebrows, nose, mouth, and other facial contours. This makes facial landmark detection act as the backbone of many applications like face verification [5], and head-pose estimations [6]. Recently facial landmark detection studies are focused on convolutional neural networks(CNNs) [7–9]. These CNN-based facial landmark detection models require a huge number of parameters and thus induce a computational overhead.

Knowledge distillation has proven to be useful in neural network parameter reduction [10, 11]. Knowledge distillation was first defined by Bucilua et al. [12], and further generalized by Hinton et al [10]. It is essentially "an effective technique that has been widely used to transfer information from one network to another network whilst training constructively" [13]. Knowledge distillation is symbolized by its ST framework, where the model providing knowledge is called teacher, and the one receiving the information(also, referred to as Dark Knowledge) is called student. This proposed work would attempt to distill knowledge from larger and more sophisticated network, to train compact student network to achieve better performance for facial landmark detection.

2 Related Work

The majority of work associated with knowledge distillation has been historically done for solving classification tasks [13], but facial landmark estimation is largely formulated as a regression problem [14, 15]. Further, there has been limited work on the use of knowledge distillation for landmark detection [16–18]. A closely related work on the use of knowledge distillation for compact facial landmark detection in a supervised setting is done by Lee et al. in their paper "Teacher and Student Joint Learning for Compact Facial Landmark Detection Network" [16], where they present a joint learning framework. During the training, the teacher and student share a convolution network to extract facial component features, then each digresses to separate regression networks, and the distillation loss is calculated between the responses of these separate regression networks, as depicted in figure 2. Post-training, the compact student network(composed of the common backbone

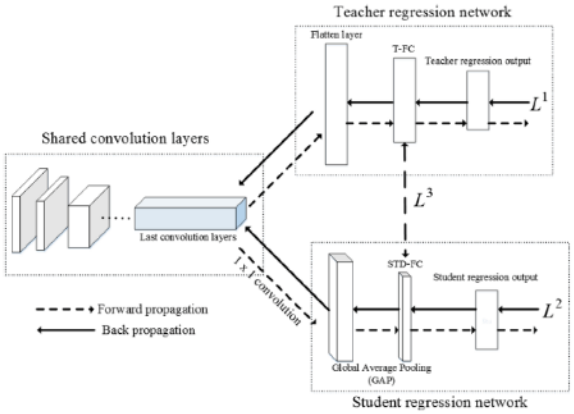

Figure 1: Learning Scheme from Lee et.al [16]

and compact regression network) proved to yield better performance. But, it can be observed that in the proposed learning scheme by Lee et.al [16], both the teacher and student share the same convolutional learning block for training, whereas the knowledge distillation happens amongst the fully-connected layer. This makes the learning scheme closely resemble a Siamese learning scheme, which causes both student and teacher to have the same number of trainable parameters, and size for the convolutional block. This work will show that a much more student compact network(both convolutional and fully-connected blocks) can achieve a comparable performance of a teacher with a much larger convolutional block, through the proposed distillation technique of this work.

Another related work is presented by Dong et.al in their work titled "Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection" [17], where the authors present an ensemble learning approach. Dong et.al propose a learning scheme(depicted in figure 2) where two students learn from a larger teacher network, and post-training students present their result as an ensemble of their responses. Although Dong et.al present compact student networks, but they need to ensemble results from both the students either parallelly or sequentially. This would contribute to an additional processing or computational overhead as compared to a solution with a single compact student network of the same size. This work presents a distillation technique that produces a single student network with high performance, to minimize the computational overhead suited for edge computing.

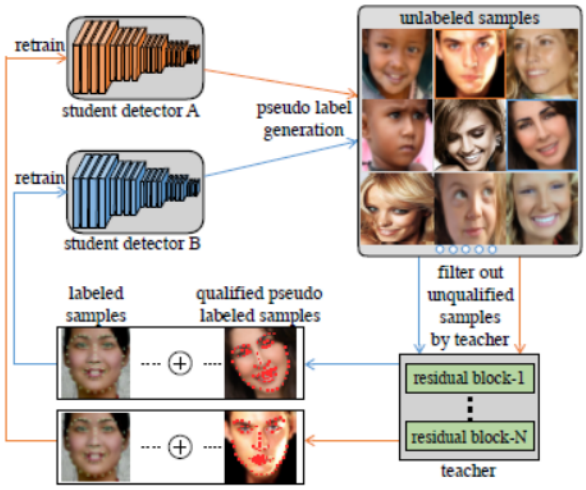

Figure 2: Learning Scheme from Dong et.al [17]

3 Overview

3.1 Data

This study would use 300W dataset [19] for training and testing the networks. This dataset was chosen as this dataset incorporates four popular datasets: LFPW [20], AFW [4], HELEN [21], XM2VTS [22], and more 135 images in extreme and difficult head poses. For purpose of making networks more robust, we augmented data by random jittering, and random rotation within 10 degrees to produce 6666 training images, and 1008 test images. Another important reason of choosing 300W is that it allow comparing the performance of trained networks against other state-of-art networks, as 300W has been used as a benchmark in the literature.

3.2 Network Architecture

It is essential to select an architecture that is suited for 68 facial landmark detection. The testbed student network architecture should be a lightweight CNN version, similar to TCDNN [15]. On the other hand, the teacher network should be a more sophisticated network resembling the work proposed by Merget et al. [23] which has proven to be more robust. Thus, this study uses teacher(ResNet50) and student(Stud5) as depicted in table 1. Their respective number of trainable parameters are listed in table 2. It can be noted that the number of trainable parameters is much smaller(84%) than the trainable parameters of teacher. These selections are further justified in the following subsection about teacher and student respectively.

Net Block	ResNet50	Stud5
conv1	$\left[7 \times 7, 64\right] \times 1$	$\left[3 \times 3, 64\right] \times 1$
conv2	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\left[3 \times 3, 128\right] \times 1$
conv3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\left[3 \times 3, 256\right] \times 1$
conv4	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\left[3 \times 3, 512\right] \times 1$
conv5	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\left[3 \times 3, 512\right] \times 1$
fc	[2048, 136(68x2)]	[512, 136(68x2)]

Table 1: Student and Teacher Network. Convolution layer presented as [kernel size, output channels] and fully connected presented as [input channels, output channels]

Model	Trainable Parameters
ResNet50(Our Teacher)	23,780,424
Stud5(Our Stud)	3,982,344

Table 2: Trainable Parameters in teacher and student

3.2.1 Teacher Network

ResNets has proved its performance in facial landmark detection in literature [24]. The original authors introduced several structures for ResNets [25]. They are different in terms of the number of layers, the number of convolutional layers in each residual block, and the filter sizes in each layer. According to the depth of the network, ResNet can be classified as ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152. The units of the networks can be classified as basic units and bottleneck units. A basic block is a skip-connection block that learns residual functions regarding the layer inputs, instead of learning unreferenced functions. A bottleneck block includes 1x1 convolutions to create a bottleneck to reduce the parameters and matrix multiplications. Due to limited local computational resources, and for purpose of rapid prototyping it was decided to settle for mid-sized ResNet, namely ResNet50. Further, after training on 300W, it outperforms many state-of-art landmark detection models as seen in table 3.

Model	NME
ResNet50(Our Teacher)	1.53
3DDE [26]	3.13
DCFE [27]	3.24
CHR2C [28]	3.3
CNN-CRF [29]	3.30

Table 3: Normalized Mean Error(NME) of teacher against state-of-art models on 300W

3.2.2 Student Network

The student network architecture had to be a lightweight CNN version, similar to TCDNN [15]. A similar network(Stud5) to TCDNN was designed, incorporating inspiration from VGG [30]. As depicted in table 1 the network had 5 convolution blocks, each with 1 convolutional layer with out channels as 64, 128, 256, 512, and 512 respectively. Each convolutional layer is followed by a max-pool layer(kernel 2x2), and finally, the output of convolutional blocks flows into a fully connected block. Further, this student network was also chosen as it has comparable performance to TCDNN of similar size when trained on 300W dataset, as detailed in table 4.

Model	MSE
Stud5(Our Student)	4.33
TCDNN [15]	4.80

Table 4: Mean Squared Error(MSE) of student against TCDNN on 300W

3.3 Knowledge Distillation

The major contribution of this work is Convolutional Knowledge Distillation(CKD) for facial landmark detection as detailed in section 3.3.2. For purpose of contrasting a classical Response Based Distillation(RBD) is also presented.

3.3.1 Response Based Distillation(RBD)

Hinton et.al [10] introduced a framework(depicted in figure 3) for knowledge distillation, where we optimise networks to minimise loss L_{RBD}, which is given as,

$$L_{RBD} = L(y_{true}, P_S) + \lambda L(P_T, P_S) \tag{1}$$

Here, L is loss criterion for network optimization, y_{true} is true label, P_S is student prediction, P_T is teacher prediction, and λ is the tuning parameter. The original paper

by Hinton et.al [10] was focused on classification and hence used L as cross-entropy. But our target problem of facial landmark detection is a regression problem [14, 15], thus we have to adapt equation 1 to suit a regression training scheme. Literature [17, 31] largely suggest the use of L as the euclidean norm of the difference of responses. The use of the euclidean norm can be largely attributed to ease-of-use and smoothness of the euclidean norm. Thus, in our experiments, we used remodelled equation 1 as equation 2, which suits our regression type problem.

$$L_{RBD} = ||y_{true} - P_S||_2^2 + \lambda||P_T - P_S||_2^2 \tag{2}$$

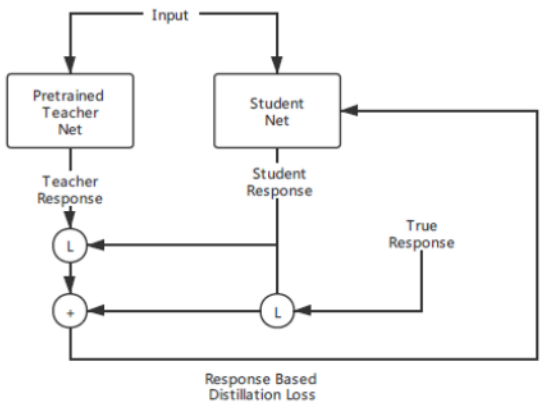

Figure 3: Response Based Distillation

3.3.2 Convolutional Knowledge Distillation(CKD)

The RBD style of KD is targeted to distill(or regularise) the student network based on the response difference between student and teacher. However, we propose that "dark knowledge" of the intermediate but more discriminatory convolutional layers of teacher can be distilled into the convolutions of the student to improve its performance. This idea is inspired by the extension of curriculum learning proposed by Bengio et.al [32], where they propose that by using "guidance hints on an intermediate layer during the training, one could considerably ease training". However, they use pre-determined heuristics, largely for classification tasks. Given that the teacher is wider and deeper than the student, distilling knowledge at any intermediate layer would need the use of a regressor. As we move towards the end of a network, the last layer has culmination of knowledge from the previous layers. Thus, ideally we should distill knowledge from the last layers. But in our case, we avoid the distillation from the last fully-connected layer because of 2 reasons. Firstly, both our network ends at a single fully connected layer of different

input size transforming input into response, thus adding a regressor for knowledge distillation from student to the teacher would not be feasible. Secondly, if we had more fully-connected layers at the end, adding a regressor would dramatically increase the parameters and hence the memory consumption. Hence we propose distilling the knowledge from the last convolution layer of the teacher(ResNet50) to the last convolution layer of the student(Stud5) by transforming the outputs from the teacher using a regressor to feed knowledge to a narrower student layer. Further, the training has to be done in 2 phases(depicted in figure 4):

- **Phase 1**: Adding a temporary convolutional regressor at the end, and dropping the fully connected for the student network, and train to optimize on the loss given by,

$$L_{CKD} = ||F_T(x, W_T) - R(F_S(x, W_S), W_R)||_2^2 \tag{3}$$

  Here, x in the training input, F_T and F_S are the nested deep function till the last convolutional layer of the teacher network and student network respectively, W_T and W_S are the weights associated to the respective last convolutional layers of the teacher and student network respectively, R and W_R are the introduced convolutional regressor and its associated weights. The convolutional regressor has to be selected such that it can transform the output of the student's last convolution layer to match the output of the teacher's last convolution layer.

- **Phase 2**: Drop the convolutional regressor from the student, reattach the fully connected layer, and retrain on the training data to adjust the fully-connected layer weights.

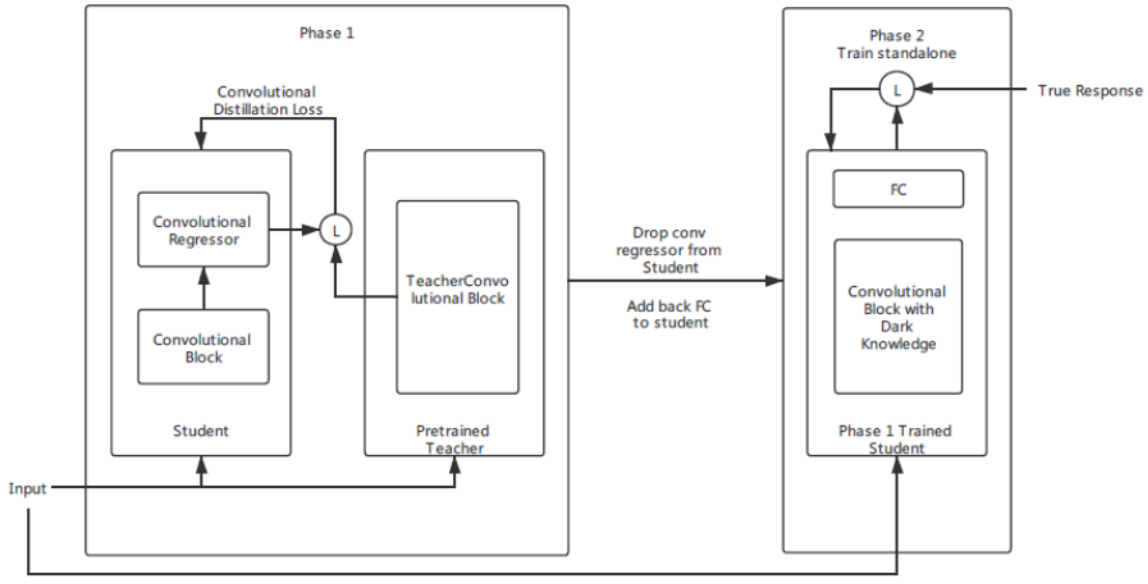

Figure 4: Convolutional Distillation Loss

4 Experiments

4.1 Design

The entire experiment was divided into training the networks on our augmented training set of 300W with the following configurations:

- ResNet50Base: Training ResNet50(Teacher) architecture standalone with mean-squared error(MSE) criterion.
- Stud5Base: Training Stud5(Student) architecture standalone with MSE criterion.
- Stud5RBD: Training Stud5 architecture with Response Based Distillation(RBD) from ResNet50Base
- Stud5CKD: Training Stud5 architecture with Convolutional Knowledge Distillation(CKD) from ResNet50Base.

4.2 Results

The above listed configurations after training were tested on the test subset of 300W and the results are summarised in table 5. It can be observed that our proposed Convolutional Knowledge Distillation(CKD) has led to a compact student Stud5CKD, which outperforms a much larger teacher for 68 facial landmark detection.

Configuration	Normalised Mean Error	Mean Squared Error
ResNet50Base	1.53	1.58
Stud5Base	2.60	4.33
Stud5RBD	2.51	4.06
Stud5CKD	**1.40**	**1.33**

Table 5: Results of experiment

5 Conclusion

This work proposed a more compact and accurate deep network for 68 facial landmark detection by distilling knowledge from a more robust and larger network using the proposed distillation criterion, namely Convolutional Knowledge Distillation(CKD). The experiments provide empirical evidence that the dark knowledge from a larger network can be distilled through the last convolutional block to boost the performance of a much compact(and faster) student network to match/outperform the larger(and slower) teacher. Specifically, if we consider the performance on 300W dataset, the compact student network trained with CKD criterion has outperformed the teacher network and other state-of-art methods listed in this paper.

References

[1] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[2] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[3] David Cristinacce, Timothy F Cootes, et al. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006.

[4] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[5] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[6] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11):3195–3208, 2007.

[7] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5861–5870, 2020.

[8] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M Robertson, and Jinqiao Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3467–3476, 2019.

[9] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[12] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[13] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[14] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[15] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.

[16] Hong Joo Lee, Wissam J Baddar, Hak Gu Kim, Seong Tae Kim, and Yong Man Ro. Teacher and student joint learning for compact facial landmark detection network. In *International Conference on Multimedia Modeling*, pages 493–504. Springer, 2018.

[17] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2019.

[18] Rongye Meng, Sanping Zhou, Xingyu Wan, Mengliu Li, and Jinjun Wang. Teacher-student asynchronous learning with multi-source consistency for facial landmark detection. *arXiv preprint arXiv:2012.06711*, 2020.

[19] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[20] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[21] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[22] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[23] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 781–790, 2018.

[24] Fine-grained facial landmark detection exploiting intermediate feature representations. *Computer Vision and Image Understanding*, 200:103036, 2020.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[26] Roberto Valle, José M Buenaposada, Antonio Valdés, and Luis Baumela. Face alignment using a 3d deeply-initialized ensemble of regression trees. *Computer Vision and Image Understanding*, 189:102846, 2019.

[27] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[28] Roberto Valle, José M Buenaposada, and Luis Baumela. Cascade of encoder-decoder cnns with learned coordinates regressor for robust facial landmarks detection. *Pattern Recognition Letters*, 136:326–332, 2020.

[29] Lisha Chen, Hui Su, and Qiang Ji. Deep structured prediction for facial landmark detection. *Advances in Neural Information Processing Systems*, 32:2450–2460, 2019.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[31] Chong Wang, Xipeng Lan, and Yangang Zhang. Model distillation with knowledge transfer from face classification to alignment and verification. *arXiv preprint arXiv:1709.02929*, 2017.

[32] Çağlar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *The Journal of Machine Learning Research*, 17(1):226–257, 2016.