Towards interpreting deep learning models for industry 4.0 with gated mixture of experts

Alaaeddine Chaoub, Christophe Cerisara, Alexandre Voisin, Benoît Iung

Towards interpreting deep learning models for industry 4.0 with gated mixture of experts

1st Alaaeddine Chaoub
Universite de Lorraine
LORIA
Nancy, France
alaaeddine.chaoub@loria.fr

2nd Christophe Cerisara
Universite de Lorraine
CNRS, LORIA
Nancy, France
cerisara@loria.fr

3rd Alexandre Voisin
Universite de Lorraine
CNRS, CRAN
Nancy, France
alexandre.voisin@univ-lorraine.fr

4th Benoît Iung
Universite de Lorraine
CNRS, CRAN
Nancy, France
benoit.iung@univ-lorraine.fr

***Abstract*—In this work, we propose to use the Gated Mixture of Experts to interpret a deep learning model trained on industrial data. Unlike monolithic deep learning models, gated modular neural networks enable to decompose parts of the models in a way that may potentially be interpreted by domain experts or users. We first propose to transform a model that gives state-of-the-art performances on a standard industrial benchmark according to this paradigm. The model aims at predicting the remaining useful life of an asset in the field of prognostics and health management for industry 4.0. Then, we experimentally validate that the performances of the transformed model are not degraded and that the resulting model segments and clusters the data streams according to an emerging concept that reflects previously published analyses by experts on this dataset, even though such a concept has never been introduced at training time. This work thus confirms the interpretable properties of the Gated Mixture of Experts in a new domain. We further study some potential weaknesses of this paradigm, in particular the excessive variability of the resulting decomposition across experiments, and we propose to modify the loss with a new knowledge-based constraint term that encodes a known prior distribution of latent concepts in the data. We show that this term enables greater control over the Gated Mixture of Experts that results in a decomposition of significantly better quality on our benchmark.**

***Index Terms*—Interpretability, Gated mixture of experts, Prognostic, Deep learning, LSTM, CMAPSS**

I. Introduction

The modular gated neural network (GMNN) is an architecture that consists of a set of individual neural network modules without shared parameters and a gated neural network that acts as a soft switch to determine which module will be used for each data sample. This approach has demonstrated multiple potential advantages such as enabling transfer learning [1], leveraging domain knowledge [2], and facilitating parallelization and distributed computing [3].

Most research works that include the gating mixture of expert have mainly focused on its overall performance and rarely on its interpretability potential. Unlike monolithic neural networks, this approach is potentially inherently interpretable since the gating networks may select modules in a way that domain experts or users could understand. A few papers investigate these capabilities but their studies focus on classification problems or on cases where the modular gated neural network is used as a whole model.

This work contributes to this area of research by investigating the potential of gated mixture of experts (GMoE)'s inherent interpretability in the context of a regression task and when the GMoE is applied to a sub part of the model. We explore whether the vanilla GMoE architecture can decompose the task in an interpretable way, and we introduce and investigate a way to incorporate human knowledge (*i.e.* through a prior distribution) into the approach.

In both cases, we provide a detailed analysis of what the gating network learns, showing that this approach can indeed produce an interpretable but not perfect decomposition, and also, we show that the proposed way of integrating human knowledge into the approach could significantly improve the quality of the task decomposition.

The rest of this paper is structured as follows. Section 2 introduces related works on interpretability and gated mixture of experts. Section 3 describes the GMoE approaches for interpretability. Section 4 highlights the results of the proposed approaches in terms of interpretability and predictive performance. Finally, conclusions and discussion are provided in section 5.

II. Related work

Due to ethical and legal requirements, interpretability is becoming increasingly crucial in deep learning models. [4] have conducted a comprehensive review, providing a 3D taxonomy for a better understanding of the distribution of research papers in this area. (a) Dimension 1 divides the approaches into passive interpretation, also known as post hoc explainability, vs. active, where the developers actively change the network architecture or training process for more interpretability. (b) Dimension 2 classifies them based on the type of explanations, for example by attribution where credits are assigned to input features, or by hidden semantics where we try to make sense of certain hidden neurons/layers/modules. (c) Dimension 3 splits the methods based on the ability to interpret the decision logic for all samples (Global), for a group of samples (Semi-local), or for individual samples (Local).

This work is part of the project AI-PROFICIENT which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957391. The Grid5000 computing resources have been used to partly train and evaluate the proposed models.

GMNN or GMoE is an approach that has been around for three decades [5] that is based on the fact that dividing a task into appropriate subtasks seems to make it easier for users/humans to understand and debug. Following the taxonomy provided by the previous review, GMoE can be viewed as an active approach where the users try to explain what the model learns and its predictions by unveiling part of the hidden semantics. The GMoE may also be viewed as enabling global interpretability when the task decomposition is perfect, but so far, according to literature results, it can be considered as semi-local only as it often fails to decompose the task perfectly.

To the best of our knowledge, few articles investigate the interpretability potential of the GMoE approach. [6] stacked two GMNNs into a single architecture to predict the class label on a randomly translated version of the MNIST dataset; their results show that this approach can learn to develop location-dependent experts at the first layer, and class-specific experts at the second layer.

Using a toy example with a 2D 6-classes Gaussian mixture, [7] show that this approach may in some cases produce an interpretable task decomposition; to confirm these results, they did further experiments on a modified version of the MNIST and FMNIST dataset, and to some extent, the model learns to allocate tasks among experts in an interpretable way, but this allocation remains unpredictable as it varies from one experiment to another.

Interpreting the model might be particularly important for more complex and realistic tasks. Hence, [8] used a GMoE model with a sparse gating mechanism in a medical use case; by embedding and visually analysing the output of this gating network, they were able to aid interpretation of patient subtype separation. In another use case, by checking the agreement or disagreement between individual experts outputs, [9] used the GMoE approach to gain insights into decision making process for semantic segmentation.

In this work, we are also interested in making the deep learning models more interpretable in a new concrete use case related to a regression task for industry 4.0.

III. GMoE approaches for task decomposition

The gated mixture of experts is a system of m experts $o_i(\cdot)$ with $i \in \{1, ..., m\}$ and a gating network $g(\cdot)$. Every expert processes the same input vector x but returns a different output vector $o_i(x)$. Typically, the gating network computes the posterior $p(i|x)$ from the same input x with a softmax:

$$g(x) = [p(\,1\,|\,x\,), \ldots, p(\,m\,|\,x\,)]$$

The final output of the system is computed as in (1):

$$f(x) = \sum_{i=1}^{m} p(\,i\,|\,x\,) \times o_i(\,x\,) \tag{1}$$

This GMoE may be used as a model of its own, or it may be inserted as a sublayer in a larger neural network [10]. In this work, we adopt this approach and insert the GMoE inside a larger state-of-the-art neural network that processes industrial time series to predict the end of life of an equipment. More precisely, we replace the preprocessing layer of this large model by a GMoE. Intuitively, the preprocessing component aims at removing from the raw sensor streams the sources of variability that are irrelevant for the target task, i.e., predicting the end of life. We thus expect the GMoE to decompose these sources of variability into interpretable clusters for the given data.

A. Simple GMoE

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be a training dataset consisting of n observations with p features, and $y_1, \ldots, y_n \in \mathbb{R}$ the corresponding gold values to predict (e.g., the remaining life time). The simple GMoE model $\hat{f}$ is trained with empirical risk minimization to maximize its performances with regard to the prediction objective:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}(f(x_j), y_j) \tag{2}$$

where $\mathcal{F}$ is some model family and $\mathcal{L}$ a loss function.

B. GMoE with constraints based on domain experts knowledge

In related works [7], [10], [11], and as also shown in the results section below, relying on the gated network alone often leads to the problem of low diversity in experts usage, as the gating network tends to converge to a state where it always produces large weights for the same few experts. This imbalance is self reinforcing: when favored experts are trained rapidly in the beginning of training, they will be more and more selected by the gating network, and thus the gap will increase resulting in poor interpretability.

Multiple approaches have been introduced to solve this imbalance [12], [13]. In this work, we choose to address this issue by integrating basic knowledge (frequency distribution of concepts present in the data) as a form of constraint to force diversity in the use of experts.

The idea is to add a posterior regularization term to the loss function that encourages the frequency distribution of Experts $\{p(\cdot|x_j)\}_{1 \leq j \leq n}$ to match a known prior. Inspired by [14], we use a MSE loss for this additional term. The new loss function $\mathcal{L}'$ is defined by (3):

$$\mathcal{L}' = \mathcal{L}(f(x), y) + \lambda' * MSE(\hat{\Omega}, \Omega) \tag{3}$$

Where $\hat{\Omega}$ represent the frequency distribution of experts, Ω represents the prior distribution of tasks, and λ' is a scalar hyperparameter controlling the strength of the constraint.

We compute an end-to-end differentiable frequency distribution $\hat{\Omega}$ of experts in two steps: first, normalizing the gated network logits with a soft-max with low temperature approximates a one-hot vector where the dominant expert (with the highest probability) has a value close to 1 and the others have values close to 0. Second, we sum and normalize these one hot vectors across the batch to get $\hat{\Omega}$. The temperature has been arbitrarily set to 0.001.

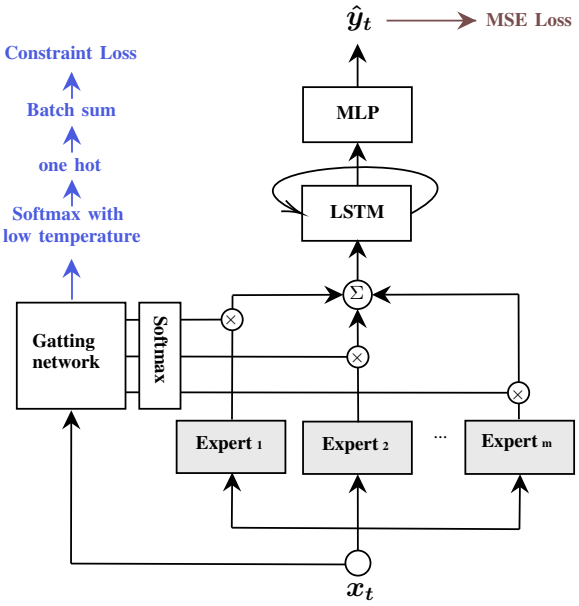

Fig. 1. GMoE-LSTM-MLP architecture with m experts.

IV. Case study: Interpreting prognostic model

A. Experimental setup

The C-MAPSS dataset [15] is a widely used benchmark in the literature for evaluating approaches for remaining useful life (RUL) prediction of turbofan engines, i.e., the remaining time before it breaks. It is divided into four sub-data sets (FD001 through FD004). In this work, our objective is to analyze whether a deep learning model trained to predict the RUL also learns to decompose the problem according to operating conditions. An operating condition (OC) can be defined as the circumstances under which an equipment operates, different operating conditions may lead to different sensor values. Therefore, we focus our experiments on the FD002 dataset, which contains six operating conditions and one failure mode (one possible cause of the engine failure).

The data contains 24-dimensional time series that correspond to measurements of sensors equipping a simulated turbofan. 3 of these 24 inputs represent the operating conditions of the turbofans: because we are interested in recovering the operating conditions, these 3 sensors are not included as input to our model, but are rather used to compute the gold/ground truth cluster label in our experiments, which is named OC.

Because in this dataset every time series lasts until the engine breaks, the true RUL is a simple decreasing linear function: the remaining time until the end of the series. However, following standard practices [16]–[18], the engine is considered to be in an healthy state as long as more than 130 timesteps of useful life remains. This gold RUL is used to train our regression model, while the ground truth OC are not used to train the model, but only to compute the clustering evaluation metrics. In other words, our approach is supervised with respect to the RUL but unsupervised with respect to the operating conditions. For model development, a set of 260 operating-to-failure trajectories is provided. We use 75% of the trajectories as the training subset, and 25% as the validation subset. In the test set, 259 sequences are provided to test the performance of the proposed approaches.

Among the state of the art approaches for this dataset, [19] proposed a model that outperforms other approaches when multiple OC are present due to its design architecture which is an end-to-end trained MLP-LSTM-MLP. In their paper, they observed that the first MLP is able to reduce the sources of variability that are not relevant for RUL prediction. The OCs in the FD002 dataset are considered one of these sources because they can change from sample to sample, leading to measurement values with different averages while the health of the turbofan engine smoothly degrades.

We propose to replace this first MLP stage by a GMoE in order to interpret the clusters that the GMoE creates. The resulting model thus follows a GMoE-LSTM-MLP architecture as shown in Fig. 1: the experts and the gating network are MLPs, as in the architecture of the original model. Indeed, this architecture has been selected since it is expected that the first part of the model, *i.e.* GMoE, will be able to retreive/discover that the turbofan measurement were done under several operating conditions (OC).

In the following, we use the same hyperparameters as those proposed in [19], while duplicating the same first MLP architecture for the experts and the gating network.

Our gated network (GN) outputs at each timestep a distribution over the experts; the argmax of this distribution is the predicted cluster. Both gold and predicted clusterings may be compared with the normalized mutual information, which is defined as follows [20]:

$$NMI(OC, GN) = \frac{2 \times I(OC, GN)}{[H(OC) + H(GN)]} \text{ , With:}$$
$$I(OC, GN) = H(OC) - H(OC \,|\, GN) \tag{4}$$

where H is the entropy, and I is the mutual information between both clusterings. The NMI is an external measure between 0 (no mutual information/ independent clusterings) and 1 (perfect correlation/ same clusterings).

B. Results and discussion

We know from [15] that 6 OCs occur in this dataset. In a real-world scenario, the number of OCs may not be known, for instance when the data are post-processed. In addition, they are usually estimated by experts, which is subject to error. We experiment next with 6 and 9 experts to assess the robustness of our approach to an erroneous prior about the number of discrete conditions. Furthermore, because of random initialization, the results may vary across different training runs, we thus run each experiment 20 times to compute the variance. Model parameters are chosen on the validation corpus by manually testing a few reasonable values. In particular, early stopping is used during training, with a maximum number of epochs set to 2000. The model with the

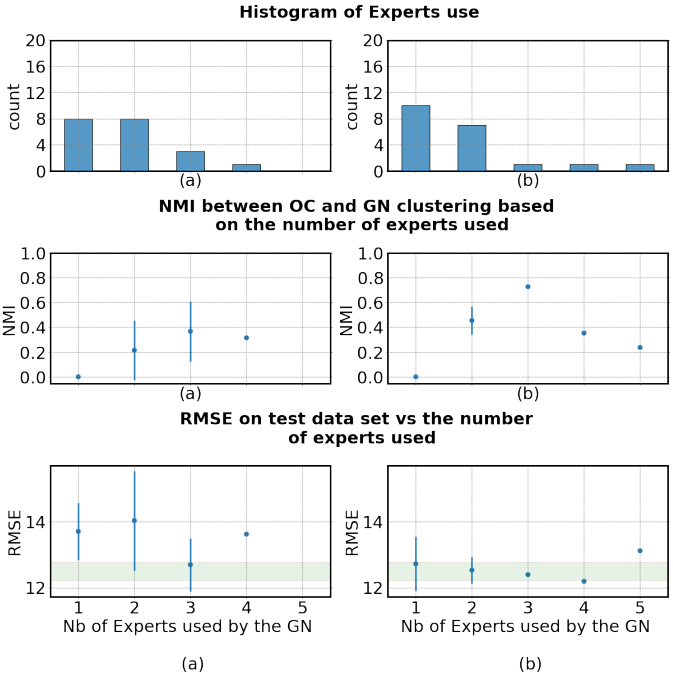

Fig. 2. Clustering evaluation when the simple GMoE-LSTM-MLP is trained on all data; left column (a): $m = 6$; right column (b): $m = 9$

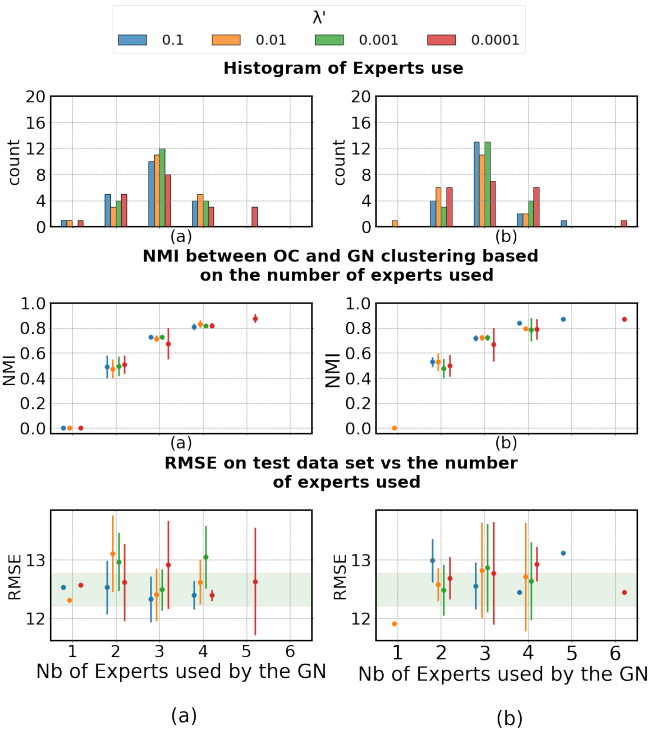

Fig. 3. GMoE-LSTM-MLP with knowledge based constraint results, the constraint value is not part of the validation loss; left column (a): $m = 6$; right column (b): $m = 9$

lowest validation loss is selected for evaluation on the test corpus.

In the following plots, the X-axis represents the number of experts actually used by the gated network, or in other words, the number of clusters predicted by our gated network $g(\cdot)$. Indeed, the gated network computes a posterior over the experts $g_i(x) = p(i|x)$ with $1 \leq i \leq m$, and we can thus associate to every input x a dominant expert $\arg\max_i g_i(x)$. The number of clusters N_c is thus the total number of experts that are dominant over the whole corpus:

$$N_c = |\{\arg\max_i g_i(x_j)\}_{1 \leq j \leq n}|$$

Every plot is structured in 3 rows and 2 columns:

- A maximum of $m = 6$ (resp. $m = 9$) experts is set in the left (resp. right) column.
- The top row shows the histograms of the number of predicted clusters N_c over the 20 experimental runs realized.
- The middle row shows the mean and standard deviation of the NMI between the predicted clusters and the operating conditions.
- The bottom row shows the root mean square prediction error values (RMSE) on the test data; for comparison, a green rectangle presenting the mean and standard deviation of the state-of-the-art RMSE from [19] is also shown.

1) Simple GMoE results: Fig. 2 evaluates the clustering produced by the simple GMoE-LSTM-MLP. In more than 87% of all runs, only one or two experts are actually used, and in these cases, the corresponding NMI is at most 0.5. Also, when 9 experts are available to the GMoE (right column), it may use up to $N_c = 5$, but these cases are rare (5% of the runs). These results show that although the gating network fails to recover the 6 expected clusters, when the numer of clusters N_c increases until 3, the resulting clusters are more and more correlated to the operating conditions (the NMI reaches 0.5, and 0.7 in one run); then, for larger N_c up to 5, the NMI decreases, which suggests that the target 6 clusters are too specific for the simple GMoE. Regarding the predictive performance of the model, the RMSE is often slightly worse than the performances reported in the state-of-the-art, except in one case. Also, we remark that the best model (with the lowest validation loss) is always found around epoch 500 regardless of the number of predicted clusters.

2) GMoE with knowledge-based constraint results: We assume a uniform prior frequency distribution Ω with the constrained loss in (3). This loss is then used to train the model parameters to optimize the main objective, RUL prediction, while inciting the model to decompose the process into 6 independent experts according to the prior Ω. However, to select the best model with early stopping, we use the unconstrained loss in (2), because this is the most important supervised task objective. Experimental results when also using (3) for hyper-parameter tuning are given later.

Fig. 3 shows the clustering quality and performances of the models in this case. We observe that mostly 3 clusters are predicted regardless of the number of experts used in the architecture, they better match the OCs with an NMI = 0.7. We also note that as the number of experts used increases,

the NMI also increases in a logarithmic way, unlike in our first experiments without any constraint. Hence, training with this constraint encourages the model to decompose the data in an interpretable way even more when the number of experts used is high. Furthermore, the number of predicted clusters is not larger than the number of OCs even when encouraging the model to use all 9 experts (Fig. 3 (b)). As for the model predictive performances, we remark that they do not vary much across conditions, and the RMSE values are very close to those of the state of the art represented by the green rectangle. Also, we note that changing the strength of the constraint in this case does not result in a significant change in model interpretability or performance.

The best models in this setup are selected around epoch 500 regardless of the constraint strength, similarly as with the simple GMoE. This shows that this approach leads to better interpretability without impacting the quality of the model.

We further investigate what happens when the final model is both trained and selected based on the RUL prediction and the knowledge-based constraint, i.e., with (3). As expected, these experimental conditions result in better interpretability: indeed, we observe that mostly 4 clusters are predicted with an NMI around 0.8. However, the predictive performances of the model significantly degrade in most conditions. The best epoch for early stopping is also much more variable than before, which suggests that the constraint term makes the convergence of training more sensitive to the random initial conditions. Furthermore, it seems that both the prediction and constraint terms are optimized successively among epochs, and in such a case, it is best to select the model based on the main target loss, even though the constraint loss may not yet have reached its optimum. Therefore, we suggest to only use the constraint for parameters training, and not for hyper-parameters tuning.

3) Comparison with related works: Table I shows that the prediction performances of the state-of-the-art model do not degrade when modified with the constrained gated mixture of experts. This suggests that it is possible in this context to design deep learning models that are more interpretable than others and still give state-of-the-art performances.

TABLE I
PERFORMANCE COMPARISON WITH RELATED METHODS: RMSE ON THE C-MAPSS FD002 DATASET. STANDARD DEVIATIONS ARE GIVEN WHEN AVAILABLE.

HDNN [17]	15.24
CapsNet [18]	16.30 $\pm$ 0.23
MLP-LSTM-MLP [19]	12.49 $\pm$ 0.28
GMoE-LSTM-MLP (m = 6) with constraint	12.59 $\pm$ 0.54
GMoE-LSTM-MLP (m = 9) with constraint	12.72 $\pm$ 0.62

V. CONCLUSION

We propose in this work an interpretable model based on the Gated Mixture of Experts for predicting the remaining useful life of engines in industrial applications. We show that both state-of-the-art prediction performances and interpretable clusters may be obtained on a standard prognostic benchmark when exploiting a prior knowledge about the distribution of the operating conditions of the engine. We analyze multiple versions of the gated mixture of experts and study the variability and sensitivity to random initial conditions of the prediction results and convergence of the clustering process. In future works, we plan to extend this approach to capture other major factors that influence the degradation process, in particular the degradation mode, and evaluate the model on other datasets and industry-related tasks, such as anomaly detection.

REFERENCES

[1] M. S. Dobre and A. Lascarides, "Combining a Mixture of Experts with Transfer Learning in Complex Games," in *AAAI Spring Symposia*, 2017.

[2] M. F. Pradier, J. Zazo, S. Parbhoo, R. H. Perlis, M. Zazzi, and F. Doshi-Velez, "Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible," *CoRR*, 2021.

[3] M. Ryabinin and A. Gusev, "Towards crowdsourced training of large neural networks using decentralized mixture-of-experts," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[4] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A Survey on Neural Network Interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, 2021.

[5] R. A. Jacobs, M. I. Jordan, S. E. Nowlan, and G. E. Hinton, "Adaptive mixture of experts," 1991.

[6] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *CoRR*, 2014.

[7] Y. Krishnamurthy and C. Watkins, "Interpretability in gated modular neural networks," in *eXplainable AI approaches for debugging and diagnosis.*, 2021.

[8] Z. Huo, L. Zhang, R. Khera, S. Huang, X. Qian, Z. Wang, and B. J. Mortazavi, "Sparse gated mixture-of-experts to separate and interpret patient heterogeneity in ehr data," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021.

[9] S. Pavlitskaya, C. Hubschneider, M. Weber, R. Moritz, F. Hüger, P. Schlicht, and J. M. Zöllner, "Using mixture of expert models to gain insights into semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[10] L. Kirsch, J. Kunze, and D. Barber, "Modular networks: Learning to decompose neural computation," *Advances in Neural Information Processing Systems*, vol. 2018-December, no. NeurIPS, 2018.

[11] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. ICLR*, 2017.

[12] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *CoRR*, 2017.

[13] E. Bengio, P. Bacon, J. Pineau, and D. Precup, "Conditional computation in neural networks for faster models," *CoRR*, vol. abs/1511.06297, 2015.

[14] T. Kim, J. Oh, N. Kim, S. Cho, and S. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," *CoRR*, vol. abs/2105.08919, 2021.

[15] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *international conference on prognostics and health management*. IEEE, 2008.

[16] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in *international conference on prognostics and health management*. IEEE, 2017.

[17] A. Al-Dulaimi, S. Zabihi, A. Asif, and A. Mohammadi, "Hybrid deep neural network model for remaining useful life estimation," in *International Conference on Acoustics Speech and Signal Processing*. IEEE, 2019.

[18] A. R.-T. Palazuelos, E. L. Droguett, and R. Pascual, "A novel deep capsule neural network for remaining useful life estimation," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 234, no. 1, 2020.

[19] A. Chaoub, A. Voisin, C. Cerisara, and B. Iung, "Learning representations with end-to-end models for improved remaining useful life prognostic," in *European Conference of the Prognostics and Health Management Society*, 2021.

[20] T. O. Kvalseth, "Entropy and correlation: Some comments," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17, no. 3, 1987.