Real-time Multi-Lingual Hate and Offensive Speech Detection in Social networks using Meta-Learning

Deepak Prasad
Computer Science and Engineering
National Institute of Technology
Warangal, India
prasad_951945@student.nitw.ac.in

K.V. Kadambari[1]
Computer Science and Engineering
National Institute of Technology
Warangal, India
kadambari@nitw.ac.in

Raghav Mukati
Computer Science and Engineering
National Institute of Technology
Warangal, India
mukati_951955@student.nitw.ac.in

Sunny Singariya
Computer Science and Engineering
National Institute of Technology
Warangal, India
singar_951949@student.nitw.ac.in

***Abstract*—A rapid increase in users on social media has given rise to a vast amount of user-generated content, including hate speech and offensive language. Such content can have serious negative consequences, ranging from psychological harm to inciting violence and discrimination. Existing studies have explored different deep learning and Natural language processing (NLP) methods to perform hate speech detection, and these solutions have yielded significant performance. Most existing solutions are limited to detecting hate speech only in English with less focus on content generated in other languages, particularly in low-resource or regional languages. The goal of this paper is to address this challenge of hate speech detection for low-resource languages and propose a tool that could provide a real-time prediction for social media posts. In this study, the main focus was on English, Hindi, Hinglish, Bengali, and Marathi languages which are commonly used in social media platforms in India. A meta-learning-based model was employed to perform hate speech detection in these languages. The proposed method helps to overcome the limitation of data scarcity and provides fast adaptation to an unseen target language. Extensive experiments were conducted on datasets comprised of different regional languages spoken in India. Accuracy, Precision, recall, and F1-score metrics are used to evaluate the model's performance. The results show that when the dataset size is small, meta-learning-based models perform better than traditional fine-tuned language models.**

***Index Terms*—Offensive language, Hate speech, meta-learning**

I. Introduction

The presence of hateful content in public spaces and social media platforms leads to the disruption of public harmony. The hate-inducing content disrupts peace in online social communities and negative influence on society as a whole. Therefore, identifying and keeping a check on hate speech in online social media is an important issue. This justifies that specialized attention is needed for various regional contents in India.

India is a land of many languages. Since the majority of existing solutions [1], [2] and [3] are focused on the English language there is a need for models that can detect hate speech in other regional languages. Detection and removal of hateful content will help in providing a better experience for users to socialize and share their opinions and also overcome concerns such as cyberbullying, abusive behavior, hate speech, and racial and sexual discrimination. Recent efforts have been put in this direction by works such as [22], and [23] however most of them are dedicated to only high-resource languages and detect hate speech majorly for English alone. Identifying hatred and offense in regional languages such as Hindi or Hindi-English mixed, Telugu, Bengali, or other Indian languages is a challenging task. The lack of sufficient annotated data in regional languages and the lack of existing models to detect hate speech in these native languages is the main motivation behind this project.

To tackle the challenges of low-resource hate speech detection, a meta-training strategy was implemented. Firstly, the model was trained with a high-resource language like English which allows it to learn a general representation of what constitutes hate speech. Then, meta-learning was employed to train this model on several comparable low-resource languages, including the target language. Since the model already has a strong initial point and it can adjust more efficiently with minor modifications to its learned representation for a given low-resource language.

II. Related Work

A. Hate Speech and Offensive Language Detection

The online social media environment is increasingly being contaminated by abusive language, which can take the form of hateful and offensive expressions, cyberbullying, discrimination, racism, sexism, misogyny, and more. Consequently, numerous studies have focused on creating automated methods to detect such content types on social media platforms. Most studies have concentrated on developing ML models for hate speech and offensive language identification in a single language - English. These studies have relied on simple

[1]Corresponding author : kadambari@nitw.ac.in

feature engineering techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) [4], Bag-of-Words (BoW), along with various traditional supervised classifiers like Support Vector Machines [7], Naïve Bayes classifier [6] and Random Forest [5].

Due to the advancements in neural network models and the abundance of labeled data primarily available in English, hate speech detection has seen the utilization of several neural network-based approaches. However, it is important to acknowledge that most of the existing research and applications in this domain have predominantly focused on English, limiting their applicability to other languages. These approaches encompass a range of models, including Recurrent Neural Networks (RNNs) [8], Long Short-Term Memory (LSTM) networks [11], Convolutional Neural Networks (CNNs) [9], bidirectional LSTMs (BiLSTMs) [1], and Gated Recurrent Units (GRUs). These models have been effectively employed in hate speech detection, leveraging their capabilities to analyze sequential and contextual information, extract relevant features, and classify instances of hate speech.

B. Data Bootstrapping based Abusive Language Detection

While a lot of work has primarily been monolingual for hate speech detection, some recent studies have addressed the challenge of offensive speech detection for low-resource languages. A recent study proposed data bootstrapping approaches [13] to detect hate speech in different Indic languages. The approach involves generating synthetic or augmented data to overcome labeled data scarcity by translation-based bootstrapping and rule-based bootstrapping.

The various bootstrapping techniques demonstrate how synthetic data generation and rule-based detection can be utilized to expand limited labeled datasets and improve a model's effectiveness in low-resource settings. The majority of such contributions used pre-trained transformer-based models: Bidirectional Encoder Representations from Transformers (BERT) [14], Multilingual Bidirectional Encoder Representations from Transformers (mBERT) [14], Robustly Optimized BERT Pretraining Approach (RoBERTa) [15], A Lite BERT (ALBERT) [16], etc. with fine tuning and data-augmentation strategies to tackle the problem of offensive language detection.

C. Ensemble Based Hate Speech Detection

A recent research study presented an ensemble-based approach for hate speech detection in the Hinglish language. The authors utilized various models from the scikit-learn library, including Logistic Regression, Random Forest, and Support Vector Machines (SVM) [7] to create an ensemble model for classification. The voting classifier was employed with the voting type set to 'hard' for the final result of the ensemble model [11]. The model achieved promising performance in detecting hate speech in Hinglish.

D. Meta Learning

Meta-learning often described as "learning to learn" is a subset of machine-learning techniques that can be leveraged to classify input with limited availability of labeled data for a target task. Meta-learning has been used in few-shot and zero-shot settings for machine translation [25] and for offensive language detection in cross-lingual [26] and code-mixed texts [27]. Meta-learning can be performed using three approaches broadly: optimization-based [29], metric-based [30] and model-based [28]. In this paper we explore MAML, an optimization-based technique for offensive speech detection in social networks.

III. Datasets

The datasets used for this project comprise the major languages spoken in India. Each of these datasets contains examples of normal and abusive speech written by people on different social media platforms. A brief description of the datasets is provided below and also in Table I. The numbers corresponding to each language in Table I represent the number of sentences of each class present for that language.

- **English** : A large fraction of offensive speech datasets available on web are in the English language. Among these, we selected the public Twitter dataset by Davidson et al for our project [17].
- **Hindi** : The dataset used for Hindi is written in Devanagari Hindi which consists of around 4.5k tweets taken from Twitter and Facebook [18].
- **Hindi-Mixed (Hinglish)** : Hinglish is a common language used in social media in India. An enormous fraction of Hindi-speaking people in India writes in Hindi-mixed which are Hindi words written using English characters [19].
- **Bengali** : Bengali is spoken by a significant population in the eastern and northeastern parts of our country. The dataset was created by crawling Facebook Posts and YouTube comments [21].
- **Marathi** : For Marathi, posts from Twitter were crawled by [20] to generate a dataset.

TABLE I
Dataset description

Language	Dataset domain	Abusive	Normal	Total
English [17]	Twitter	6898	17,607	24,505
Hindi [18]	Twitter & Facebook	1433	3161	4594
Hinglish [19]	Twitter	3438	4841	8279
Bengali [21]	Facebook & Youtube	3298	6702	10,000
Marathi [20]	Twitter	876	1623	2499

IV. Proposed Methodology

The proposed solution in this paper aims to solve the problem of offensive speech detection in low-resource languages with limited labeled training data. It achieves this through the use of the meta-learning technique and a few-shot cross-lingual approach. This model can quickly adapt to an unseen target language with only a small number of labeled examples, making it an effective tool in the fight against online hate speech.

The use of meta-learning, also known as "learning to learn" has emerged as an effective technique for solving few-shot learning problems, including offensive speech detection in low-resource languages. The scarcity of datasets in regional or low-resource languages poses a significant challenge for offensive speech detection. However, this issue can be tackled by using meta-learning which enables fast adaptation to an unseen target language with only a small number of

labeled examples [12].
Meta-learning has already shown impressive performance in various computer vision tasks, including the classification of new image classes with limited examples of that class. In the context of offensive speech detection, meta-learning enables the model to learn from a limited amount of labeled data and generalize to new languages quickly [10].

A. Model Agnostic Meta-Learning (MAML)

MAML is a model-agnostic approach to meta-learning, which can be applied to a wide range of machine-learning models without requiring modifications to their architecture. The MAML algorithm involves two phases: the inner loop and the outer loop. In the inner loop, the model is trained on a small number of examples from a task to update its parameters. In the outer loop, the updated parameters are used to evaluate the model's performance on a separate set of examples from the same task. The gradients obtained from the outer loop are then used to update the shared initialization, which can be used to adapt to new tasks quickly.
The parameters are updated through gradient descent in the following manner:

$$\theta_i^{'} \leftarrow \theta - \beta \nabla L_{T_i}(f_\theta) \tag{1}$$

where β and L denote step size and loss value respectively. The meta-learner optimization aims to minimize the meta loss computed from the training tasks

$$\min_\theta \sum_{i=1}^{m} L_{T_i}(f_\theta) = \sum_{i=1}^{m} L_{T_i}(f_\theta - \beta \nabla_\theta L_{T_i}(f_\theta)) \tag{2}$$

The parameter θ is updated to

$$\theta = \theta - \gamma \nabla_\theta L_{T_i}(f_{\theta_i'}) \tag{3}$$

where γ is the meta-learner learning rate. Model Agnostic Meta-Learning (MAML) is a powerful approach to meta-learning that has shown promising results in a variety of domains, like computer vision, and Natural language processing (NLP). Its model-agnostic nature makes it a versatile algorithm that can be applied to a wide range of machine-learning models.

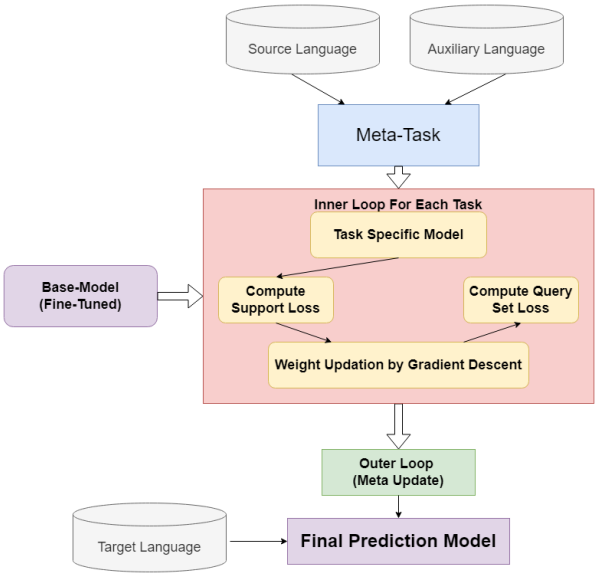


Fig. 1. Proposed Model Flow

The model comprises two primary components: the base model and the meta-learner. The base model is responsible for learning the input data's representation, tokenizing the input sentences, and making predictions for the target language. On the other hand, the meta-learner is responsible for learning the adaptation rules and adjusting the base model's parameters to enhance its performance on the target language. In the proposed methodology, the model is first trained on a support set, which consists of examples from the target language along with similar auxiliary language after which the model makes predictions on the query set, which consists only of unlabeled examples of the target language. This training ensures that the model reaches a good parameter initialization before getting fine-tuned on the target language. The support and query sets are used to compute loss and update the model parameters accordingly. This model has demonstrated encouraging results in detecting hate speech across multiple languages. For the solution, Cross-lingual Language Model - RoBERTa (XLM-R) [24] is being used as the base model. XLM-R is a state-of-the-art multilingual transformer-based encoder developed by Facebook AI. It is trained on multiple languages and has shown impressive performance on a range of natural language processing tasks.

The proposed model can be used to deploy a real-time hate speech detection tool for social media networks. An API that can be integrated with the various social media platforms would provide real-time predictions on these posts and help in classifying them as hateful or non-hateful. Since the meta-learning-based models can provide prediction in zero-shot settings i.e. even when there is an absence of labeled examples for a language, the above solution could prove viable and effective in enhancing online safety and promoting responsible communication.

V. Experiments and Results

A. Results and Discussions

The performance of the proposed model was evaluated on each of the low-resource languages stated above. For each of the models, the batchSize was set to 8, inner loop learning to 2e-5, and outer loop learning rate to 1e-5. The models were trained for 15 epochs and then evaluated based on accuracy, precision, and recall. The meta-learning approach outperforms the fine-tuned models in almost all the metrics. Moreover, for languages having small datasets such as Marathi the performance difference was even more as can be seen from the table below.

TABLE II
META-LEARNING MODEL RESULTS

Language	Accuracy	Precision	Recall
Hindi	0.80	0.69	0.65
Hindi-mixed	0.83	0.78	0.81
Bengali	0.88	0.80	0.84
Marathi	0.86	0.78	0.82

The results in Table II show that meta-models perform with an accuracy of more than 80% for all the low-resource languages with equally good recall and precision values with the exception of Hindi. Although the dataset size for some languages such as Marathi is as small as 2.5k tweets the proposed model shows a strong performance. This testifies to the effectiveness of meta-learning models when there is data scarcity.

To gain a deeper insight into the proposed models we compare it with the state-of-the-art fine-tuned models based on transfer learning. The bert-based Cross-lingual Language Model (XLM-R) was fine-tuned on the available datasets and compared with the corresponding meta-learning-based models.

TABLE III
FINE TUNING VS META-LEARNING (MAML)

Language	Fine tuning			Meta Learning		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Hindi	0.79	0.72	0.55	0.80	0.69	0.65
Hinglish	0.83	0.80	0.78	0.83	0.78	0.81
Bengali	0.87	0.80	0.82	0.88	0.80	0.84
Marathi	0.81	0.76	0.62	0.86	0.78	0.82

For all the languages, the meta-learning model outperforms the fine-tuning approach in all the metrics especially recall which is an important metric for classification problems. The superiority of meta-learning models was also testified by comparing model performance for different dataset sizes. For our experiments, we have used Bengali language for comparison where we gradually vary dataset size from 2000 to 10000.

Fig. 2. F1-score comparison for different dataset sizes(Bengali)

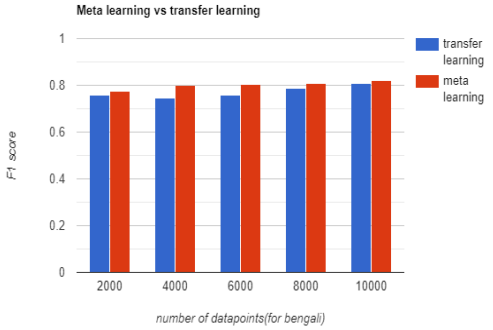
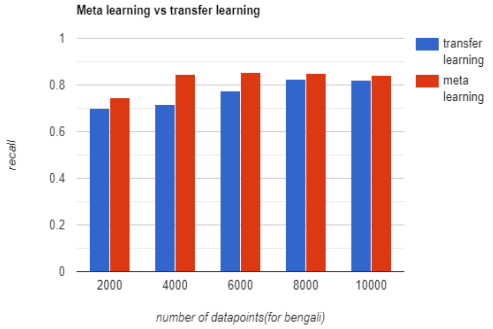


Fig. 3. Recall comparison for different dataset sizes(Bengali)



Experiments with different dataset sizes show how the proposed model outperforms the fine-tuned models especially for small dataset sizes or in cases of data scarcity. It can be observed from Fig. 2 and Fig. 3 that difference in performance between the two approaches is much more in small dataset sizes.

VI. Conclusion and Future Work

The experiments and results in this study show that the meta-learning approach can be used for developing hate speech recognition even in the case of data scarcity. When data is sufficiently available, the meta-learning-based models show similar performance as fine-tuned models. This suggests that with sufficient data available we can choose any methodology and achieve similar performance. In this study, four low-resource languages spoken in India have been picked and attempted to create hate-speech detection models. Undertaking a similar course we can develop models to support other low-resource languages spoken in India such as Tamil, Telugu, etc. The proposed approach can be extended to solve other problems where data scarcity is a hurdle in training large transformer models. A multi-class hate speech classification model can be developed to classify tweets as normal, offensive, and hateful. Analysis of different language families to understand how they play a role in the overall understanding of the model and affect the performance when included as auxiliary language can help in improving low-resource language models. Future improvements can be done to detect hateful content in images, audio files, and videos.

REFERENCES

[1] A. Kumar, V. Tyagi and S. Das, "Detection of Offensive Language in Social Networks Using LSTM and BERT Model," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), Arad, Romania, 2021, pp. 546-548, doi: 10.1109/ICCCA52192.2021.9666342.

[2] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in Proc. 1st Workshop Abusive Lang. Online, 2017, pp. 41–45.

[3] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, 2018, doi: 10.1145/3232676

[4] Y. Mehdad and J. Tetreault, "Do characters abuse more than words?" in Proc. 17th Annu. Meeting Special Interest Group Discourse Dia logue, Los Angeles, CA, USA, 2016, pp. 299–303. [Online]. Available: https://www.aclweb.org/anthology/W16-3638

[5] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on twitter," in Proc. NAACL Student Res. Workshop, San Diego, CA, USA, Jun. 2016, pp. 88–93. [Online]. Available: https://www.aclweb.org/anthology/N16-2013

[6] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," J. Exp. Theor. Artif. Intell., vol. 30, no. 2, pp. 187–202, Mar. 2018, doi: 10.1080/0952813X.2017.1409284.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proc. Int. AAAI Conf. Web Social Media, 2017, pp. 512–515. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in Proc. 10th ACM Conf. Web Sci., New York, NY, USA, Jun. 2019, pp. 105–114, doi: 10.1145/3292522.3326028.

[9] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion), Geneva, Switzerland, 2017, pp. 759–760, doi: 10.1145/3041021.3054223.

[10] L. Gao and R. Huang, "Detecting online hate speech using context aware models," in Proc. Recent Adv. Natural Lang. Process. Meet Deep Learn. (RANLP), Varna, Bulgaria, Nov. 2017, pp. 260–266, doi: 10.26615/978-954-452-049-6_036.

[11] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)," in Proc. 14th Workshop Semantic Eval., Barcelona, Spain, 2020, pp. 1425–1447. [Online]. Available: https://aclanthology.org/2020.semeval-1.188

[12] S. Thrun and L. Pratt, Learning to Learn: Introduction and Overview. Boston, MA, USA: Springer, 1998, pp. 3–17, doi: 10.1007/978-1-4615-5529-2_1.

[13] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages. In Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22). Association for Computing Machinery, New York, NY, USA, 32–42. https://doi.org/10.1145/3511095.3531277

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. Minneapolis, MN, USA, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692

[16] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A dictionary-based approach to racism detection in Dutch social media," 2016, arXiv:1608.08738.

[17] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 11

[18] Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Forum for Information Retrieval Evaluation. 29–32.

[19] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media. 36–41

[20] Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 437–443.

[21] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In Proceedings of International Joint Conference on Advances in Computational Intelligence. Springer, 457–468

[22] Vashistha, N.; Zubiaga, A. Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media. Information 2021, 12, 5. https://doi.org/10.3390/info12010005

[23] Mohiyaddeen, Mr. and Siddiqi, Sifatullah, Automatic Hate Speech Detection: A Literature Review (July 15, 2021). Available at SSRN: https://ssrn.com/abstract=3887383 or http://dx.doi.org/10.2139/ssrn.3887383

[24] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 8440–8451.

[25] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 3622–3631.

[26] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-lingual few-shot hate speech and offensive language detection using meta learning," IEEE Access, vol. 10, pp. 14880–14896, 2022.

[27] G. V. Suresh, B. R. Chakravarthi, and J. P. McCrae, "Meta-learning for offensive language detection in code-mixed texts," in Proc. Forum Inf. Retr. Eval., Dec. 2021, pp. 58–66

[28] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in Proc. 33rd Int. Conf. Mach. Learn. (Proceedings of Machine Learning Research), vol. 48, M. F. Balcan and K. Q. Weinberger, Eds. New York, NY, USA: PMLR, 20–22, Jun. 2016, pp. 1842–1850. [Online]. Available: http://proceedings.mlr.press/v48/santoro16.html

[29] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. 34th Int. Conf. Mach. Learn. (Proceedings of Machine Learning Research), vol. 70, D. Precup and Y. W. Teh, Eds. Sydney, NSW, Australia, Aug. 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html

[30] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1– 11. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf