Attention-based connectivity matrix for autism spectrum disorder diagnosis from fMRI data*

Sirine Sboui[1,2], Mohamed Djallel Dilmi[1,*] and Faten Chaieb-Chakchouk[1]

[1]Efrei Paris Panthéon Assas University, 30-32 Avenue de la République, 94800-Villejuif, France
[2]Macif, 64 RUE RENE BOULANGER, 75010-PARIS, France

Abstract

This paper introduces a novel approach for autism spectrum disorder (ASD) classification from resting-state fMRI by replacing conventional Pearson correlation connectivity matrices with attention-based representations. From a theoretical standpoint, we demonstrate that Pearson correlation is a special case of the scaled dot-product attention mechanism under fixed projections and normalization, which allows attention to be viewed as a strict generalization of functional connectivity. Building on this formulation, we propose an end-to-end pipeline in which attention layers directly learn connectivity patterns from fMRI time series, eliminating the need for predefined functional correlation matrices. Experiments on the ABIDE-II dataset show that attention-based connectivity significantly improves diagnostic performance across diverse architectures, including MLP, vanilla Transformers, BrainNetCNN, and the Brain Network Transformer (BNT). The proposed method achieves a state-of-the-art accuracy of 90.4% with BNT, substantially outperforming correlation-based baselines. These results highlight the potential of attention mechanisms to enhance both the accuracy and interpretability of functional connectivity modeling in neuroimaging.

Keywords

Attention mechanism, functional connectivity matrix, Pearson correlation, Representations Fusion, fMRI data, spatiotemporal data, multivariate timeseries, Autism spectrum disorder,

1. Introduction

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental condition that disrupts brain functioning, particularly in areas related to social interaction, communication, and behavior [1]. Symptoms typically emerge in early childhood, but their type and severity can vary widely among individuals. The diagnosis of autism generally relies on clinical assessments based on behavioral observations and standardized questionnaires completed by caregivers.

A thorough medical history and examination may also be necessary to rule out other conditions with similar symptoms. Despite being commonly used, these diagnostic methods are often subjective and time-consuming, which can lead to delays in identification. Such delays may hinder timely access to specialized treatments that are crucial for improving quality of life and fostering social inclusion. To address these limitations, non-invasive neuroimaging techniques have emerged as valuable tools for diagnosing neurological and mental disorders, including ASD. These techniques, encompassing structural and functional modalities, provide critical insights into brain anatomy and connectivity. Functional magnetic resonance imaging (fMRI), one of the most widely used functional neuroimaging methods, measures cerebral activity by detecting changes in cerebral blood flow, which are closely linked to oxygenation levels across different brain regions. One common approach to construct a brain network in the neuroimaging community is via pairwise correlations between BOLD time courses from two regions of interest (ROIs) [2] [3]. To obtain reliable and usable brain signals, rigorous preprocessing is essential. This preprocessing pipeline typically includes a series of corrections (such as for motion and slice acquisition timing), spatial normalization to a reference space, temporal filtering, and spatial

Cooperative Information Systems - Early Research Achievement and Demos, 2025.
*Corresponding author.
sirine.sboui@externe.efrei.fr (S. Sboui); djallel.dilmi@efrei.fr (M. D. Dilmi); faten.chakchouk@efrei.fr (F. Chaieb-Chakchouk)
0000-0002-7909-290X (M. D. Dilmi); 0000-0002-2968-2426 (F. Chaieb-Chakchouk)

smoothing.

To enhance diagnostic accuracy, these data are increasingly leveraged by artificial intelligence (AI) tools. Machine learning (ML) and deep learning (DL) have proven effective in extracting relevant patterns from complex neuroimaging data.

Attention mechanisms [4] are attracting growing interest in neuroimaging and are being explored in depth by the scientific community. Initially developed for sequential data processing in automatic natural language processing (NLP), these mechanisms are now being adapted to brain data modeling. Based on the assumption that functional connectivity conveys sufficient discriminative information to differentiate autistic from non-autistic subjects, integrating an attention mechanism enables extracting finer, more refined representations. In addition, this approach contributes to a better understanding of brain mechanisms by identifying the brain regions affected and their functional interactions.

2. State of the art

Autism has long been the subject of considerable scientific inquiry and remains a primary focus of contemporary research. Neuro-scientists seek to understand the impact of autism on the structure and functioning of the brain in order to understand better the mechanisms underlying this disorder. Neuroimaging techniques, including positron emission tomography (PET), diffusion tensor imaging (DTI), and magnetic resonance imaging (MRI), have been extensively used in autism research, particularly because they provide rich structural and functional information. Building on the wealth of information provided by these neuroimaging techniques, functional MRI (fMRI) data can be analyzed by studying functional connectivity (FC), notably using correlation matrices such as Pearson's correlation, partial correlation, or tangent correlation, which model relationships between different regions of the brain. Pearson's correlation matrices, for example, have been widely used in autism research as a fundamental input for different machine learning and deep learning methods. For multi-center data, [5] uses 1001 male subjects, aged between 5 and 40 years, to evaluate traditional models such as linear SVM, RBF SVM, XGBoost, TabNet, and MLP. Among these, the SVM classifiers outperformed the others, achieving an AUC of approximately 75%. In [6], a dense neural network (DNN) classifier is used for autism classification, with an average accuracy of 88%. More advanced models, such as CNN (Convolutional Neural Networks), have been applied to this matrix type, treating it as a 2D image [7]. In addition, other approaches have exploited connectivity matrices, especially those based on Pearson correlation, using them with graph-based techniques such as graph neural networks (GNN); these methods model brain regions as nodes and their functional connections as weighted edges. As a pioneer work, [8] introduced the use of GNNs in functional connectivity networks to classify ASD. Another distinct approach based on the attention mechanism has been developed in recent years. This approach has demonstrated its effectiveness in analyzing functional connectivity (FC) for autism diagnosis; for example, the Brain Network Transformer (BNT) [9] achieved approximately 72% accuracy.

Recent research also seeks to exploit other types of correlation and connectivity matrices as inputs to deep learning methods. For instance, [10] using three types of connectivity matrices as inputs to deep learning methods, with 884 subjects, achieving an accuracy of 73.43% using cross-validation. Similarly, our work aims to replace the classical Pearson correlation matrix with another type of data based on the attention mechanism to reduce data loss, as detailed in the paragraphs below.

3. Proposed Method

Our approach consists in replacing traditional correlation matrices, typically computed during the pre-processing stage, with attention matrices learned from fMRI time series. Although correlation matrices, especially Pearson correlation matrices, are widely used in autism classification research to represent functional connectivity [11, 9], they often do not preserve the full richness of the original signals [12], and this can affect the performance of AI models. Thus, we need a more complex operator to compensate for the loss of information. In contrast, attention mechanisms are capable of directly

processing time series data while simultaneously modeling interactions between brain regions, thus eliminating the need for predefined connectivity matrices and enabling more efficient and integrated learning. From a theoretical point of view, correlation matrices can be considered as a particular case of attention matrices under specific initialization conditions. This perspective reinforces the idea that attention mechanisms generalize traditional connectivity representations while offering greater modeling flexibility and the ability to directly capture more complex relationships in brain data. To explore and validate this hypothesis, our method consists in learning the attention matrix during the training phase itself, allowing the model, typically a deep learning architecture enhanced with an attention mechanism layer, to optimize connectivity representations in a task-driven and fully end-to-end manner. Figure 1 provides a schematic overview of the architecture underlying our proposed approach.

Formally, let $X =\in \mathbb{R}^{N\times P}$ denote the data matrix, where N is the number of time points and P the number of regions (or variables) of the brain (that is, X represents a P time series series of length N).

Scaled dot-product Attention. The scaled dot-product attention mechanism [4] computes the interaction between the P elements of X through the formula

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where $Q = (W_q X)^T, K = (W_k X)^T, and\, V = (W_v X)^T \in \mathbb{R}^{N\times d}$ represent the query, key, and value matrices, and d_k is the dimension of the key vectors. Intuitively, this mechanism learns attention weights that highlight the most relevant interactions among elements of the input.

Pearson correlation matrix On the other hand, the Pearson correlation between regional time series is expressed in matrix form as

$$C(X) = \frac{X^\top X}{N}, \tag{1}$$

which captures the linear similarity between all pairs of regions. This matrix is symmetric and fixed once computed from the data, with no learnable parameters.

Attention as a Generalization of Correlation

We state that, under simple (but restrictive) choices of the projection and normalization, the attention operator expresses the Pearson correlation matrix. We then give a corollary describing the effect of initializing an attention layer with the correlation matrix.

Proposition 1 *Let $X \in \mathbb{R}^{N\times P}$ be a data matrix with N time points and P regions. Consider the scaled dot-product attention*

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

and suppose that the following specializations hold:

1. *$V = I_P$ (the $P \times P$ identity),*
2. *$Q = K = (WX)^\top$ for some linear map $W \in \mathbb{R}^{m\times N}$,*
3. *$W^\top W = I_N$ (i.e. W defines an orthonormal projection on its row-space),*
4. *the softmax is removed (i.e. we consider the raw affinity scores),*
5. *$\sqrt{d_k} = N$.*

Then the resulting attention matrix A equals the (unnormalized) Pearson correlation matrix of X:

$$A = \frac{X^T X}{N} = C(X)$$

Proof Under the above assumptions,

$$A = \frac{QK^\top}{\sqrt{d_k}}V = \frac{(WX)^\top\big((WX)^\top\big)^\top}{\sqrt{d_k}}I_P = \frac{(WX)^\top X^\top W^\top}{\sqrt{d_k}}.$$

Using $W^\top W = I_N$ and $\sqrt{d_k} = N$, we obtain

$$A = \frac{X^\top W^\top WX}{N} = \frac{X^\top X}{N} = C(X).$$

Thus, Pearson correlation appears as a degenerate case of attention with fixed, non-learned projections.

Corollary 1 *(Initialization with correlation) Consider a neural classifier that incorporates an attention module initialized with the Pearson correlation matrix* $C(X) = \frac{X^T X}{N}$*. If the training objective* $\mathscr{L}(\theta)$ *is smooth and locally convex in a neighborhood of the initialization, and gradient descent is performed with a suitably small step size, then* $\mathscr{L}(\theta)$ *decreases monotonically during training in this neighborhood. In other words, empirical performance is* monotonically decreasing *when learning starts from correlation-based initialization.*

Remark. Although global convexity is not guaranteed for deep architectures, this analysis highlights that Pearson correlation can serve as a principled initialization for attention, bridging traditional FC measures with modern learnable representations. Empirically, we observe that such initialization accelerates training and improves final performance (cf. Table 1).

Fisher Transformation

The authors of [2] proposed using a transformation to calculate functional connectivity between brain regions. Specifically, they used the Fisher transformation [13], a well-established practice in neuroscience for stabilizing the variance of correlation coefficients, to approximate the distribution of correlation values by a normal distribution.. The transformation is defined as:

$$z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \tanh^{-1}(\rho)$$

To maintain consistency with their data processing approach, we applied the same transformation to the attention matrix A. Values resulting from the Fisher transformation that are infinite due to the correlation of a region with itself being equal to 1 have been replaced by zero by convention, as the relationship between a region and itself is not the focus of our study.

4. Experiments and discussions

4.1. Dataset

The Autism Brain Imaging Data Exchange (ABIDE), comprising ABIDE I (2012) and ABIDE II (2015), has become one of the most widely used datasets [14]. It includes both structural (sMRI) and functional (rs-fMRI) brain imaging data, along with detailed phenotypic and clinical information from individuals with ASD and matched controls, collected at multiple international sites. Compared to other available resources, ABIDE is distinguished by its large sample size, public accessibility and standardized data acquisition protocols, making it a benchmark dataset for research on brain connectivity and a key enabler for the development of AI-based diagnostic tools in autism research.

In our study, we used functional MRI data from the abide dataset [14] from 17 sites to calculate the functional connectivity matrices and attention matrices. The data went through a preprocessing phase (C-PAC) and a time series extraction step based on the Craddock 200 atlas. As the acquisition protocols

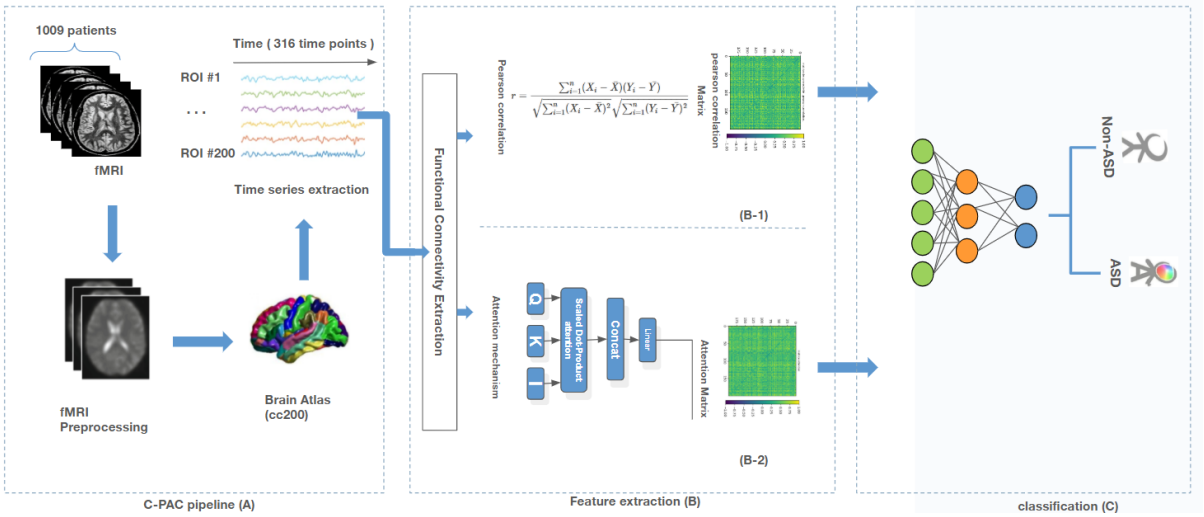

Figure 1: Schematic representation of our approach. (A) **CPAC Pipeline**: Preprocessing and denoising of fMRI data using the CPAC pipeline, including atlas application for time series extraction. (B) **Feature Extraction**: Two parallel approaches for extracting connectivity features: (B-1) *Pearson Correlation* computing classical functional connectivity matrices based on Pearson correlation coefficients; (B-2) *Attention Mechanism* learning connectivity matrices through an attention-based mechanism. (C) **Classification**: Applying a range of classifiers, from traditional machine learning models to deep learning architectures.

Table 1
Evaluation of Machine and Deep Learning Methods Using Pearson Correlation and Attention Matrices (Random vs Non-Random Split)

Models	Split	Pearson			Fusion		
		Acc	Sen	Spe	Acc	Sen	Spe
MLP	Non-Random	66.4	71.3	61.4	63.3	71.5	55.0
	Random	66.1	60.7	72.9	66.4	74.8	57.7
VanillaTF	Non-Random	68.3	69.6	67.0	71.7	71.5	72.0
	Random	65.0	70.0	57.5	88.6	85.8	91.5
BNT [9]	Non-Random	68.8	72.5	65.0	89.6	96.0	83.0
	Random	72.0	77.3	65.9	**90.4**	88.7	92.2
BrainNetCNN [7]	Non-Random	64.3	70.5	58.0	70.3	79.4	61.0
	Random	63.0	70.2	56.6	88.3	85.4	91.3

and scan times for different sites are not the same, the time series lengths are also different, generally varying between 90 and 400 timestep. we utilized a curated subset of the ABIDE dataset comprising 1009 participants, with 516 (51.14%) being Autism Spectrum Disorder (ASD) patients (positives). Each participant has resting-state fMRI time series exceeding 100 time points. This selection ensures sufficient temporal resolution for robust connectivity analyses. The data were obtained from the get-abide GitHub project, which provides an accessible and well-structured resource designed for neuroimaging research, including fMRI time series, site, labels, and Pearson correlation matrices as features.

4.2. Experimental Setup and Results Analysis

We conducted a series of experiments to validate our approach. As described above, the goal is to demonstrate the effectiveness of replacing the classical Pearson correlation matrix, commonly used in most research, with a novel matrix derived from an attention mechanism for binary classification.

To prepare our data, we applied zero-padding to standardize the lengths of the time series, setting a maximum length of 316 time points. After normalization, the dataset was split into training (70%), validation (10%), and test (20%) subsets using a stratification strategy to ensure fair and consistent

comparisons between methods evaluated in our experiments.

To evaluate model performance, we adopted two procedures. In the first, we used the Pearson correlation matrix as input data. In the second, we directly used the fMRI time series to compute attention-based representations, obtained through the fusion of an attention layer with pre-trained models.

The models used in this study include a Multilayer perceptron, a vanilla Transformer and two pretrained models: BNT [9] and BrainNetCNN [7]. Each model was assessed according to Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe).

Additionally, we employed two data splitting strategies. The first, referred to as the "Non-random" split, involved stratification by both site and label, with a fixed random seed. This approach helps control for confounding factors and class imbalance by ensuring all acquisition sites are well represented, thus reducing the risk of overfitting to site-specific features. On the other hand, we performed repeated random subsampling with 5 splits, stratifying only on labels. This method does not guarantee representation of all sites in each split, which may lead to potential data leakage if the same site appears in both training and testing subsets. To address this variability, we averaged the performance results across the 5 random splits.

Comparing the metrics obtained from both splitting methods, for MLP and both types of procedures, the overall results remain limited, likely due to the simplicity of this model, which has relatively few parameters and is unable to capture complex patterns in the data. This implies that neither the Pearson correlation matrix nor fusion is efficient for robust classification with MLP (Table 1).

MLP performance remained limited across both splitting strategies, likely due to its simplicity and low capacity to capture complex patterns. Consequently, neither Pearson correlation matrices nor fusion yielded robust classification with this model (Table 1).

We then evaluated deeper architectures. The vanilla Transformer performed better on the non-random split (68.3% vs. 65% with random), while the BNT model was more effective under random splitting (72% vs. 68.8% with Pearson features), suggesting that its pre-training captures stable patterns across shuffled data [15]. BrainNetCNN reached 64.3% on the non-random split but dropped slightly to 63.0% with random splitting. Overall, Pearson-based approaches showed only moderate performance.

To overcome this, we introduced an attention layer, enabling end-to-end learning of attention weights directly from time series. Models trained for 30 epochs with this fusion approach achieved consistent improvements across both split types (Table 1), with stronger results for the random split.

After data augmentation, the BNT model reached the best accuracy (90.4%), followed by the vanilla Transformer (88.6%) and BrainNetCNN (88.3%). The Transformer offered balanced sensitivity and specificity (85.8% / 91.5%), while BrainNetCNN achieved 85.4% / 91.3%. BNT outperformed both, with accuracy, sensitivity, and specificity of 90.4%, 88.7%, and 92.2%. These results highlight the effectiveness of the proposed fusion strategy for reliable autism classification.

5. Conclusion and Current work

In this study, we proposed a novel paradigm for modeling functional connectivity in autism research by replacing the conventional Pearson correlation matrix with an attention-based representation directly learned from fMRI time series. This formulation bridges traditional statistical connectivity measures with modern neural architectures by showing that Pearson correlation can be expressed as a degenerate case of attention under specific initialization conditions. Beyond its theoretical grounding, the method enables connectivity matrices to be optimized in a data-driven, task-specific, and end-to-end fashion. Experimental results on a curated subset of the ABIDE-II dataset confirm the effectiveness of this approach. Across multiple architectures—including MLP, vanilla Transformers, BrainNetCNN, and the Brain Network Transformer—our attention-derived connectivity consistently improved classification performance, with the BNT model achieving a state-of-the-art accuracy of 90.4%. These findings highlight the benefits of jointly learning connectivity representations with predictive models rather than relying on fixed correlation-based features. Ongoing work explores extending this approach along

several directions. First, we aim to evaluate the method on larger and more heterogeneous datasets, such as ABCD, to assess generalization across populations. Second, we plan to investigate hybrid strategies that fuse attention-based and correlation-based features to combine interpretability with expressive power. Finally, future research will address model explainability by analyzing learned attention patterns in relation to neuroscientific priors, potentially providing new insights into the altered connectivity mechanisms underlying autism spectrum disorder.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 in order to: Grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Hodges, C. Fealko, N. Soares, Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation, Transl Pediatr 9 (2020) S55–S65.

[2] K. J. Friston, Functional and effective connectivity: a review, Brain Connect 1 (2011) 13–36.

[3] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, C. F. Beckmann, Correspondence of the brain's functional architecture during activation and rest, Proc Natl Acad Sci U S A 106 (2009) 13040–13045.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] F. Mainas, B. Golosio, A. Retico, P. Oliva, Exploring autism spectrum disorder: A comparative study of traditional classifiers and deep learning classifiers to analyze functional connectivity measures from a multicenter dataset, Applied Sciences 14 (2024). URL: https://www.mdpi.com/2076-3417/14/17/7632. doi:10.3390/app14177632.

[6] F. Z. Subah, K. Deb, P. K. Dhar, T. Koshiba, A deep learning approach to predict autism spectrum disorder using multisite resting-state fmri, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/8/3636. doi:10.3390/app11083636.

[7] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, G. Hamarneh, BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment, Neuroimage 146 (2016) 1038–1049.

[8] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, D. Rueckert, Spectral graph convolutions for population-based disease prediction (2017). arXiv:1703.03020.

[9] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, C. Yang, Brain network transformer, ArXiv abs/2210.06681 (2022). URL: https://api.semanticscholar.org/CorpusID:252125388.

[10] D. Wang, X. Yang, W. Ding, Autism spectrum disorder (asd) classification with three types of correlations based on abide I data, 2025. URL: https://www.aimsciences.org/article/id/652d06cd6e7706046e317b90. doi:10.3934/mfc.2023042.

[11] M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, G. Deshpande, Functional Connectivity-Based prediction of autism on site harmonized ABIDE dataset, IEEE Trans Biomed Eng 68 (2021) 3628–3637.

[12] G. U. Yule, Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series, Journal of the Royal Statistical Society 89 (1926) 1–63. URL: http://www.jstor.org/stable/2341482.

[13] W. H. Thompson, P. Fransson, On stabilizing the variance of dynamic functional brain connectivity time series, Brain Connect 6 (2016) 735–746.

[14] C. Cameron, B. Yassine, C. Carlton, C. Francois, E. Alan, J. András, K. Budhachandra, L. John, L. Qingyang, M. Michael, Y. Chaogan, B. Pierre, The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives, Front. Neuroinform. 7 (2013).
[15] H. Zhang, M. Cissé, Y. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, ArXiv abs/1710.09412 (2017). URL: https://api.semanticscholar.org/CorpusID:3162051.