Proc. of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023)
19-21 July 2023, Tenerife, Canary Islands, Spain

2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) | 979-8-3503-2297-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICECCME57830.2023.10253245

A deep learning approach for automatic video coding of deictic gestures in children with autism

Roberta Bruschetta *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy roberta.bruschetta@irib.cnr.it

Simona Campisi *C.O.T. Cure Ortopediche Traumatologiche S.P.A. Via Ducezio 1* 98124 Messina, Italy simona.campisi@irib.cnr.it

Marilina Mastrogiuseppe *Department of Humanities University of Trieste* Trieste, Italy marilina.mastrogiuseppe@u nits.it

Elisa Leonardi *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy elisa.leonardi@irib.cnr.it

Stefania Aiello *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy stefania.aiello@irib.cnr.it

Christian Salvatore *DeepTrace Technologies S.R.L. Via Conservatorio 17 20122 Milan, Italy* salvatore@deeptracetech.com

Alessandro Venturi *DeepTrace Technologies S.R.L. Via Conservatorio 17 20122 Milan, Italy* venturi@deeptracetech.com

Elia Schiavon *DeepTrace Technologies S.R.L. Via Conservatorio 17 20122 Milan, Italy* schiavon@deeptracetech.com

Agrippina Campisi *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy agrippina.campisi@irib.cnr.it

Francesca Isabella Famà *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy francescaisabella.fama@irib.cnr.it

Cristina Carrozza *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy cristina.carrozza@irib.cnr.it

Carla Blandino *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy carla.blandino@irib.cnr.it

Flavia Marino *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy flavia.marino@irib.cnr.it

Antonio Cerasa *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy antonio.cerasa@irib.cnr.it

Olga Capirci *Institute of Cognitive Sciences and Technologies (ISTC) National Research Council of Italy (CNR)* Rome, Italy olga.capirci@istc.cnr.it

Giovanni Pioggia *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy giovanni.pioggia@irib.cnr.it

Liliana Ruta *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy liliana.ruta@irib.cnr.it

Gennaro Tartarisco *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR)* 98164 Messina, Italy gennaro.tartarisco@irib.cnr.it

***Abstract*—Autism is a heterogeneous neurodevelopmental condition characterized by impairments in socialcommunication, along with restrictive and repetitive patterns of interests and behaviors and sensory atypicalities. Early impairments in gestural communication, especially in deictic gestures, are significantly associated with autism and strong predictors of language development. Despite the implication of deictic gestures in autism has been acknowledged, it has not been sufficiently explored by artificial intelligence. To address this, the paper proposes an automatic digital coding approach based on deep learning models. By using a transformer architecture, a multi-frame modelling strategy has been implemented and applied on 37 video clips of naturalistic mother-child interactions with the aim to recognize four main deictic gestures: pointing, giving, showing and requesting. The system was trained and validated on 31 clips, internally tested on 6 clips and externally tested on 5 extra clips, using Python. Preprocessing phase involves using a 1024 feature extractor based on Densenet121 pretrained on Imagenet. Preliminary results showed respectively 100% of accuracy for training set, 80% for validation set and 67% for internal testing set. These findings suggest that the proposed system is a very promising approach for the automatic analysis of deictic gestures. In future work, we plan to validate our model on a larger number of samples to achieve higher and more reliable performances.**

Keywords—deep learning, transformer architecture, autism, deictic gestures.

I. Introduction

Autism is a group of complex neurodevelopmental conditions characterized by impairments in socialcommunication, along with restrictive and repetitive patterns of interests and behaviors and sensory atypicalities. Symptoms of Autism Spectrum Disorders (ASD) vary widely among individuals and range from mild to severe impairment [1]. A key symptom of autism is a deficit in non-verbal communication, such as gestures [2].

Research evidence has shown that early impairments in gestural communication can represent an early biomarker for autism diagnosis, particularly for siblings with a high genetic risk of developing the condition in the early stages of development [3].

Previous studies demonstrated that the development of gestures in autism follows a distinct pattern compared to typical development (TD) [4–6] as well as other clinical populations, such as language disorders or Down Syndrome (DS) [7].

In the early years of life, TD children use gestures for many functions, such as interacting, and triangulating attention [8, 9]. Among gestures, deictic gestures, such as pointing, showing, giving and requesting, play an important role in the development of language, forming the link between actions and language and helping the children communicate intentions that they are not yet able to express verbally [10, 11].

In contrast to TD ones, children with autism show, often, significant impairments in verbal communication that are not compensated by the use of gestures. Studies have shown that gestural production of autistic children is mainly characterized by ritualized actions [12] and that the overall level of gestures is reduced and impaired compared to control groups. Moreover, Maljaars et al. [13] evidenced that in children with autism there is a prevalence of gestures expressing behavioral regulation functions rather than declarative purposes.

Understanding the complexity of gestural communication is thus crucial for a better comprehension of socio-communicative impairments in children with autism and can lead to important improvements in both early detection and personalized intervention approaches [14].

The manual micro-analytical behavioral coding from video footages of naturalistic mother-child interactions is still the major method to explore quantitatively and qualitatively gestural communication in children with autism [7]. However, it is substantially time-consuming and it requires trained human coders, limitations that represent significant barriers for the development of the field [15].

The introduction of Artificial Intelligence (AI) for automated video coding, may result greatly appealing to exceed the limits of manual coding and to provide step further in the field to be translated in clinical practice. Particularly, transformer models are a popular choice for deep learning in the vision domain and have been widely adopted in recent years for various computer vision tasks such as image classification, object detection, semantic segmentation, and image generation [16]. These architectures are effective in handling large amounts of data and capturing long-range dependencies in images, making them suitable for being applied to this field.

In [17] Hofemann et al. proposed an integrated system based on combining a trajectory recognition algorithm with symbolic object data to identify the “Pointing” gesture in a multi-modal human-machine interface. However, this research area is still largely unexplored and to the best of our knowledge, there are no studies that have applied deep learning approaches on naturalistic videos for the exploration of deictic gestures in neurodevelopmental disorders.

In this work, we present an automatic digital coding approach based on a transformer architecture to recognize four main deictic gestures from naturalistic mother-child interactions video clips.

II. Material and Methods

A. Data Collection

The explorative study was conducted on a group of 6 young children between the ages of 23 and 63 months who had been clinically diagnosed with ASD according to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [1] by expert clinicians using the Autism Diagnostic Observation Scale [18]. Children with autism were recruited at the clinical facilities of the Institute for Biomedical Research and Innovation of the National Research Council of Italy (IRIB-CNR) in Messina. The study received ethical clearance by the Ethics Committees of Azienda Ospedaliera

Universitaria P. Giaccone, Palermo, Italy, approval ID 07/2022.

For data collection and processing, we employed the detailed strategy proposed by Mastrogiuseppe et al. [7]. This strategy involved recording and analyzing a 10 minute video of naturalistic mother-child interaction for each participant. During the interaction sessions, parents were asked to play spontaneously with their children using a standardized set of age-appropriate games that included objects such as a ball, nesting cups, phone, train, tea set, doll, and blanket. These objects were chosen to promote a wide range of play behaviors, from simple exploration to the highest level of symbolization. Each of the six sessions was recorded using a fixed video camera and later it was processed through a specifically developed moment-by-moment coding procedure that allowed for both quantitative and qualitative analysis of gestural production. The codification was performed using ELAN [19], a software for video analysis and annotation which allowed us to detect and label four categories of deictic gestures (see Fig. 1).

Deictic gestures are a type of nonverbal communication that refers to an object or event by directly pointing or touching it. Their meaning is dependent on the context in which they are used. Mastrogiuseppe et al. [7] proposed the classification of children's deictic gestures into four categories:

- Pointing: it involves the child using distinctly the index finger to direct attention towards a specific object, place or event;
- Showing: when the child grasps an object towards the adult to show it;
- Giving: it involves the child offering an object to the partner;
- Requesting: when the child extends the arm with the palm facing up to ask for something, usually accompanied by opening and closing the hand.

Fig. 1. Examples of the four deictic gestures

From six video processing, 37 repetitions of deictic gestures were identified resulting in 37 separated video clips to use for model development. Detected gestures are divided as follows:

- Showing: 3 repetitions;
- Requesting: 14 repetitions;
- Giving: 11 repetitions;
- Pointing: 9 repetitions.

B. *Deep learning Classifier:The Transformer Architecture*

Transformers are simple models that exploit the attention mechanism to transform input sequences into output sequences using an encoder-decoder architecture. Positional encodings are computed using sine and cosine functions and summed to input embeddings before encoder and decoder stacks in order to include information about the relative position of the elements in the sequence [20].

Both the encoder and the decoder consist of 6 identical blocks. In the encoder, each block includes a multi-head self-attention network and a fully connected feed-forward network. Instead, decoder blocks include a further multi-head attention sublayer applied to the outputs of the corresponding encoder block.

The attention mechanism proposed in [20], called "Scaled Dot-Product Attention", applies an attention function for mapping a query Q and a set of key-value (K, V) pairs (with dimension d_k and d_v respectively) to an output. The output is computed as the weighted sum of the values, where the weight of each value is calculated by the dot product between queries of each position and keys of other positions in the sentence dividing by d_k and applying a softmax function:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In particular, in self-attention layers of the encoder, all the queries, keys and values come from the output of the previous encoder and in the same way for self-attention layers of the decoder that represents an autoregressive model.

In "encoder-decoder attention" layers instead, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder.

At the end, the decoder output is converted through a linear layer followed by a softmax function to predict next token probabilities [20].

To automatically identify the four deictic gestures from selected video clips, we developed a multi-frame approach based on artificial intelligence (AI) models considering each video clip as a whole, rather than training and classifying single frames.

In more detail, as a first step, each video clip was preprocessed by resizing its frames to the input dimensions of the DenseNet121 [21] exploited for the feature extraction procedure: 128x128x3. In addition, a fixed length of 2 seconds (25 fps) was set as duration for each clip. Shorter videos were padded to 50 frames by adding empty frames at the end, while longer videos were centered on the action of interest cutting frames in excess.

Automatic feature extraction was performed using deep learning. Specifically features were extracted from a DenseNet121 pretrained on ImageNet, by removing the fully connected layer at the top and by applying a global average pooling to the output of the last convolutional block. A total of 1024 features was thus obtained for each video frame. Extracted features from frames of each video clip were employed for training a multi-frame classification model based on the transformer architecture proposed in [20]. Information about the relative positions of the frames was incorporated into the model exploiting the positional encoding. Positions of frames within videos were encoded using an Embedding layer preceding the transformer encoder and then added to the pre-computed features.

The described architecture is followed by a maximum pooling operation (GlobalMaxPooling1D layer) and a dropout layer (0.5 Dropout layer) to prevent overfitting. Model output consists of a densely-connected layer with Softmax activation function that assigns output probabilities for the four possible categories and performs classification. The resulted model includes ~4.27 millions of trainable parameters.

Details about model architecture and trainable parameters are reported in Table 1.

TABLE I. DETAILS ABOUT MODEL ARCHITECTURE

Layer type	Output shape	Parameters
Input Layer	-	0
Positional Embedding	(None, None, 1024)	51200
Transformer Enconder	(None, None, 1024)	4211716
Global Max Pooling 1D	(None, 1024)	0
Dropout	(None, 1024)	0
Dense	(None, 4)	4100
Total parameters		***4267016***
Trainable parameters		***4267016***
Non-trainable parameters		***0***

80% of available video clips were employed for model training and validation (validation split was set to 0.15), while the remaining 20% was used for internal testing. In addition, 5 video clips recorded in a different environmental setting were used as external-testing set. Details about training, validation and testing sets are reported in Table 2.

Model training and validation were performed for 50 epochs using Adam optimization and sparse categorical cross-entropy as loss function.

All the steps of video pre-processing, feature extraction, training, validation and testing of the model were performed using Python 3 with the following libraries: Tensorflow 2.9.2, Keras 2.9.0, Pandas, Numpy, ImageIO, CV2 and VisualKeras.

TABLE II. DESCRIPTION OF TRAINING, VALIDATION AND TESTING SETS

Set	Total	Showing	Requesting	Giving	Pointing
Training + Validation	31	2	12	10	7
Internal Testing	6	1	2	1	2
External Testing	5	1	-	2	2

III. RESULTS

Classification performance of training, validation and internal testing are reported in Table 3. Results indicate that the model was able to achieve an overall accuracy of 67% in classifying the video clips into four actions of interest.

A detailed analysis of the internal testing results revealed that the model was highly accurate in identifying video clips depicting the gesture of "Requesting" with 100% accuracy (2 out of 2 clips correctly classified), and similarly for the gesture of "Giving" and "Showing" with 100% accuracy (1 out of 1 clip correctly classified for each gesture). However, the model struggled in correctly identifying video clips depicting the gesture of "Pointing", with 0% accuracy (both clips were incorrectly classified as "Giving").

When the model's performance was evaluated on an external testing set, it was found that it was able to accurately classify video clips depicting the gesture of "Pointing" and "Giving" with 100% accuracy (2 out of 2 clips correctly classified for each gesture). However, the model incorrectly classified a video clip depicting the gesture of "Showing" as "Giving" (0 out of 1 clip correctly classified). There were no video clips depicting the gesture of "Requesting" in the external testing set, hence no evaluation for this gesture was possible. More detailed information on the results of internal and external testing can be found in Table 4.

TABLE III. OVERALL MODEL PERFORMANCE

Metric	Training	Validation	Internal Testing
Accuracy	100%	80%	67%
Loss	0.01	2.02	1.17

TABLE IV. DETAILS ABOUT INTERNAL AND EXTERNAL TESTING RESULTS

Set	Instance	True Label	Output Probabilities			
			Showing	**Requesting**	**Giving**	**Pointing**
Internal Testing	1	Showing	**59.74%**	4.59%	34.49%	1.19%
	2	Requesting	7.59%	**82.17%**	3.45%	6.80%
	3	Requesting	0.57%	**72.46%**	18.46%	8.51%
	4	Giving	1.20%	13.85%	**63.40%**	21.55%
	5	Pointing	0.72%	0.71%	97.83%	**0.74%**
	6	Pointing	10.95%	0.63%	85.32%	**3.10%**
	Single Class Accuracy		100%	100%	100%	0%
External Testing	1	Showing	**6.87%**	7.97%	78.27%	6.89%
	2	Giving	18.52%	0.69%	**51.52%**	29.27%
	3	Giving	12.64%	14.51%	**71.79%**	1.06%
	4	Pointing	1.15%	10.98%	24.24%	**63.62%**
	5	Pointing	0.11%	1.02%	4.94%	**93.93%**
	Single Class Accuracy		**0%**	-	100%	100%

IV. DISCUSSION

The proposed system is a powerful tool based on deep learning models that can be used to automatically identify and classify gestures from video recordings of naturalistic actions. The system was implemented using a multi-frame approach that utilizes a transformer architecture to consider each video clip as a whole, allowing to capture frames dependencies. The primary goal of this work was to apply this system for the automation of the video coding procedure proposed in [7] for the evaluation of gestural communication deficits in children with autism, with specific focus on deictic gestures. These gestures, also known as joint attention gestures, are a fundamental early biomarker of autism.

The proposed system demonstrated an accuracy of 67% in internal testing and 80% in external testing. However, it was observed that all the misclassified clips were assigned to the class “Giving” which has a larger number of instances in the training set compared to the other classes. To improve model’s performance, it is necessary to balance the number of instances among the target classes.

Additionally, to further improve model’s generalizability and reliability, it is essential to conduct a massive training phase using larger datasets. In this regard, we are currently collecting new data. However, it is worth mentioning that the recordings used in this study have some limitations such as inconsistent quality among different videos and even within the same video (e.g., different field of view, view angle, zoom, background, and camera position), the presence of other people, and movements not properly framed by the camera. The heterogeneity of data is beneficial for developing a robust model that can handle any naturalistic video. However, to achieve better results, it is highly recommended to establish a standardized protocol for data collection.

V. CONCLUSION

In this study, we applied advanced deep learning models and innovative techniques to explore the communicative and gestural abilities of children with neurodevelopmental disorders, such as autism. Therefore, this technology has the potential to be integrated into a broader multimodal AI framework to support early screening and assessment of neurodevelopmental disorders. One of the main challenges in this field is the automatic recognition of gestures, as the task requires a high level of complexity, even for humans. Very small variations in body language can completely change the meaning of a gesture. For example, a child extending his arm to pass an object or to show it can be difficult to distinguish, as the difference between the two actions is often very subtle. These variations in meaning are typically conveyed by different nuances in the motor and expressive patterns of the gesture. To address this challenge, the field of computer vision and machine learning is actively working on developing more sophisticated algorithms to improve the ability of machines in understanding human gestures. This study demonstrates the feasibility of using AI systems to recognize specific gestures from naturalistic videos. The developed framework, which includes Python code and a notebook, can be easily and flexibly used for the analysis of new video clips in the same domain or for training model parameters for new applications in the field of computer vision. This research opens new avenues of investigation that can help gain new insights and advancements in the field of neurodevelopmental disorders. By supporting early detection and improving the development of personalized treatment protocols, this technology has the potential to lead to more effective interventions for children with autism and other neurodevelopmental disorders.

ACKNOWLEDGMENT

The research has been supported by the CNR Institute of Biomedical Research and innovation of Messina under the special project founded by MISE - Accordo per l'Innovazione DM 5.3.2018 Project n. 867 - READS - Reading Early Autism Disorders Signs, Court of Auditors registration No. 323 on April 20, 2022;

Progetto "Early Start" (Accordo Quadro con l'Azienda Sanitaria Provinciale di Catania, Prot.ASP-CT n. 0001180/2021);

Roberta Bruschetta is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma;

Simona Campisi is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVIII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

REFERENCES

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, DSM-5-TR. American Psychiatric Association Publishing, 2022. doi: 10.1176/appi.books.9780890425787.

[2] A. de Marchena and I.-M. Eigsti, 'Conversational gestures in autism spectrum disorders: asynchrony but not decreased frequency', Autism Res, vol. 3, no. 6, pp. 311–322, Dec. 2010, doi: 10.1002/aur.159.

[3] L. Zwaigenbaum et al., 'Studying the emergence of autism spectrum disorders in high-risk infants: methodological and practical issues', J Autism Dev Disord, vol. 37, no. 3, pp. 466–480, Mar. 2007, doi: 10.1007/s10803-006-0179-x.

[4] L. Bartak, M. Rutter, and A. Cox, 'A Comparative Study of Infantile Autism and Specific Developmental Receptive Language Disorder: I. The Children', Br J Psychiatry, vol. 126, no. 2, pp. 127–145, Feb. 1975, doi: 10.1192/bjp.126.2.127.

[5] T. Charman, S. Baron-Cohen, J. Swettenham, G. Baird, A. Drew, and A. Cox, 'Predicting language outcome in infants with autism and pervasive developmental disorder', Int J Lang Commun Disord, vol. 38, no. 3, pp. 265–285, 2003, doi: 10.1080/136820310000104830.

[6] R. Luyster, K. Lopez, and C. Lord, 'Characterizing communicative development in children referred for autism spectrum disorders using the MacArthur-Bates Communicative Development Inventory (CDI)', J Child Lang, vol. 34, no. 3, pp. 623–654, Aug. 2007, doi: 10.1017/s0305000907008094.

[7] Marilina Mastrogiuseppe, Olga Capirci, Simone Cuva, and Paola Venuti, 'I gesti deittici nei bambini con disturbi dello spettro autistico: un'analisi quantitativa e qualitativa all'interno di interazioni spontanee madre-bambino', Sistemi intelligenti, no. 1, pp. 11–32, 2016, doi: 10.1422/83833.

[8] J. S. Bruner, 'The social context of language acquisition', Language & Communication, vol. 1, pp. 155–178, 1981, doi: 10.1016/0271-5309(81)90010-0.

[9] S. Shumway and A. M. Wetherby, 'Communicative acts of children with autism spectrum disorders in the second year of life', J Speech Lang Hear Res, vol. 52, no. 5, pp. 1139–1156, Oct. 2009, doi: 10.1044/1092-4388(2009/07-0280).

[10] O. Capirci and V. Volterra, 'Gesture and speech The emergence and development of a strong and changing partnership', Gesture, vol. 8, pp. 22–44, Apr. 2008, doi: 10.1075/gest.8.1.04cap.

[11] J. M. Iverson, O. Capirci, and M. C. Caselli, 'From communication to language in two modalities', Cognitive Development, vol. 9, no. 1, pp. 23–43, Jan. 1994, doi: 10.1016/0885-2014(94)90018-3.

[12] L. Camaioni, P. Perucchini, F. Muratori, B. Parrini, and A. Cesari, 'The communicative use of pointing in autism: developmental profile and factors related to change', Eur Psychiatry, vol. 18, no. 1, pp. 6–12, Feb. 2003, doi: 10.1016/s0924-9338(02)00013-5.

[13] J. Maljaars, I. Noens, R. Jansen, E. Scholte, and I. van Berckelaer-Onnes, 'Intentional communication in nonverbal and verbal low-functioning children with autism', Journal of Communication Disorders, vol. 44, no. 6, pp. 601–614, Nov. 2011, doi: 10.1016/j.jcomdis.2011.07.004.

[14] N. C. Capone and K. K. Mcgregor, 'Gesture Development: A Review for Clinical and Research Practices', Journal of Speech, Language, and Hearing Research, vol. 47, pp. 173–186, 2004, doi: 10.1044/1092-4388(2004/015).

[15] K. Fujiwara, Q. S. Bernhold, N. E. Dunbar, C. D. Otmar, and M. Hansia, 'Comparing Manual and Automated Coding Methods of Nonverbal Synchrony', Communication Methods and Measures, vol. 15, no. 2, pp. 103–120, Apr. 2021, doi: 10.1080/19312458.2020.1846695.

[16] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, 'Transformers in Vision: A Survey', ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[17] N. Hofemann, J. Fritsch, and G. Sagerer, Recognition of Deictic Gestures with Context. 2004, p. 341. doi: 10.1007/978-3-540-28649-3_41.

[18] A. McCrimmon and K. Rostad, 'Test Review: Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part II): Toddler Module', Journal of Psychoeducational Assessment, vol. 32, no. 1, pp. 88–92, Feb. 2014, doi: 10.1177/0734282913490916.

[19] H. Sloetjes and P. Wittenburg, Annotation by category: Elan and iso dcr. 2008.

[20] A. Vaswani et al., 'Attention Is All You Need'. arXiv, Dec. 05, 2017. Accessed: Jan. 20, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.