Supplementary Information

An Artificial Neutral Network Chip Based on Two-Dimensional Semiconductor

Shunli Ma[1]†, Tianxiang Wu[1]†, Xinyu Chen[1]†, Yin Wang[1], Hongwei Tang[1], Yuting Yao[1], Yan Wang[1], Jianan Deng[2], Jing Wan[2], Zhengzong Sun[1], Zihan Xu[3], Antoine Riaud[1], Chenjian Wu[4], Yang Chai[5], Peng Zhou[1]*, Junyan Ren[1]*, Wenzhong Bao[1]*, David Wei Zhang[1]

[1] *State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China*

[2] *State Key Laboratory of ASIC and System, School of Information Science and Technology, Fudan University, Shanghai 200433, China*

[3] *Shenzhen Sixcarbon Technology, 188 Jiangshi Road, Shenzhen 518106, China*

[4] *School of Electronic and Information Engineering, Soochow University, Suzhou, China*

[5] *Department of Applied Physics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, P. R. China*

† These authors contributed equally to this work.

* Email: pengzhou@fudan.edu.cn; junyanren@fudan.edu.cn; baowz@fudan.edu.cn

Table of contents

1. Synthesis of wafer-scale MoS_2 film and fabrication procedure of MoS_2 chips: 2

2. Physical modeling of MoS_2 TG-FETs 4

3. Level-62 SPICE model 6

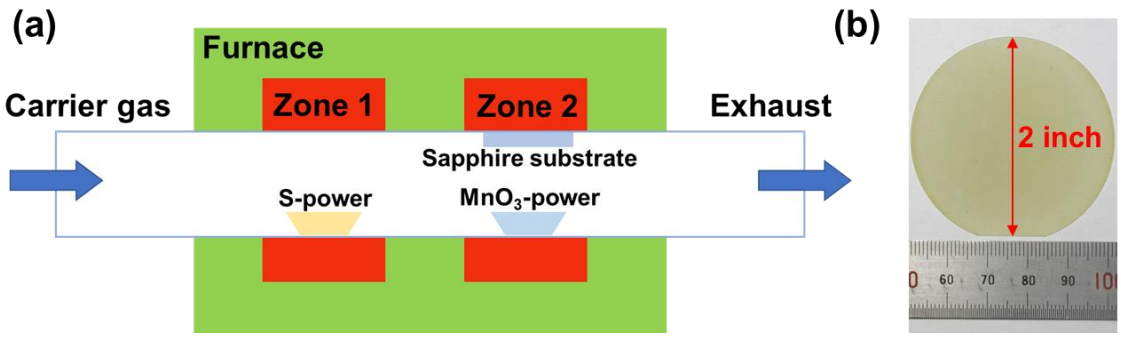
4. ANN IC structure and MAC operation analysis 10

5. Calibration: Non-ideal weight correction 14

6. Artificial neural networks and deep learning algorithm 16

1. Synthesis of wafer-scale MoS_2 film and fabrication procedure of MoS_2 chips:

Synthesis of wafer-Scale MoS_2: A crucible with MoO_3 power (Alfa Aesar 99.95%) was placed in Zone 2 with an appropriate amount of sulfur powder (Alfa Aesar 99.999%) placed in Zone 1 (upstream of the flow in the tube). The distance between the two zones was 30 cm. A carefully cleaned sapphire substrate was placed face-down on the crucible containing the MoO_3 power. During synthesis, 300 sccm of Ar was used as the carrier gas. The synthesis temperature in Zone 1 and Zone 2 was controlled at 180 °C and 650 °C, respectively. A continuous monolayer MoS_2 film was synthesized at atmospheric pressure with 10 min sulfuration time.



Supplementary Fig. 1: (a) Schematic diagram of the CVD growth process. (b) Photograph of a 2 in. sapphire wafer uniformly covered with CVD-grown MoS_2.

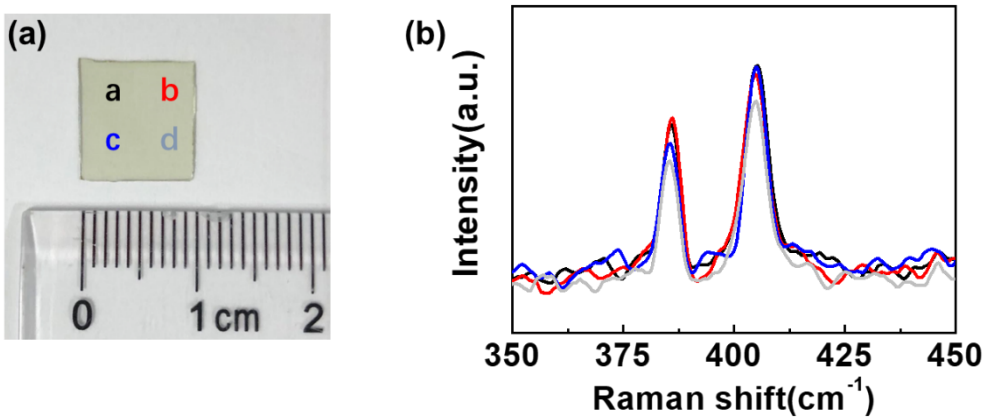
Fabrication of MoS_2 integrated circuits:

The MoS_2 FETs and circuits are fabricated on a wafer-scale sapphire substrate uniformly covered with a CVD-grown MoS_2 film. The contact electrodes (35 nm Au), source, and drain contacts were patterned with traditional laser writing tool (MicroWriter ML3) and subsequently deposited using electron beam (E-beam) evaporation. CF_4 plasma etching was performed to define the MoS_2 channel region. In order to increase V_T and reduce the leakage current in the device, a seed layer (1 nm SiO_2) was deposited using E-beam evaporation followed by furnace annealing in

oxygen at 100 °C. Then, 20-nm-thick HfO_2 was subsequently grown via atomic layer deposition (ALD) as the main dielectric layer. Then lithography is used again to define the via holes and SF_6 plasma etching was used to etch the via through the dielectric layer. Then a 30-nm-thick Au metal was deposited to fill the via holes using E-beam evaporation. A final lithography and lift-off process was used to form the top metal layer (35 nm Au), which was deposited by thermal evaporation.

Characterization and electrical measurements:

Uniformity of the CVD-grown MoS_2 film was examined using Raman spectroscopy measurements (Renishaw inVia) at four points in different regions of a 1×1 cm^2 sample. The electrical properties of the MoS_2 FETs and circuits were measured with a probe station connected to a semiconductor analyzer (Agilent B1500A).

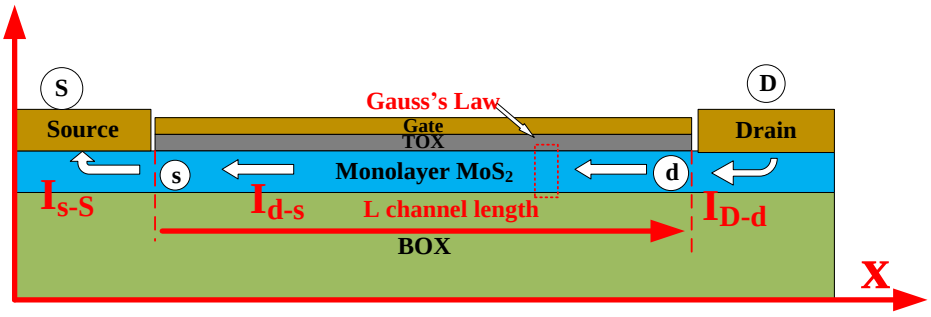


Supplementary Fig. 2 (a) Photograph of a 1×1 cm^2 MoS_2 chip cut from a 2-in. wafer. (b) Corresponding Raman spectra were measured at the points labeled a-d on the wafer.

2. Physical modeling of MoS_2 TG-FETs

The contact between the metal electrode at the source (drain) and the MoS_2 forms a Schottky contact. The device is divided into three parts: the drain output (D) to drain the contact (d), a channel region spanning from d to s, and s to the source output (S). Each section is considered a separate area with different voltage and current characteristics. Because these regions are in series, we have $\partial I_{ds}(x)/\partial x = 0$, and the following relation holds:

$$I_{D\to d} = I_{d\to s} = I_{s\to S} = I_{DS} \tag{1}$$



Supplementary Fig. 3 Schematic device structure and current distribution.

Considering interfacial defects as acceptors, the effective energy E_{it} is located below the conduction band, and the effective trap density is D_{it}. To simplify the model, it is assumed here that D_{it} is a delta function of energy, and this method can be generalized. With a certain deviation, the number of captured carriers (N_{it}) is given by:

$$N_{it} = \int_{-E_0}^{E_0} D_{it} f(E) dE = D_{it}/(1 + \exp(\frac{E_0 - E_{it} - qV_F}{k_B T})) \tag{2}$$

$$n_{2D} = N_{it} + \frac{1}{qT_{2D}}\left(\varepsilon_{TOX}\frac{V_G - \Delta\Phi_m/q - \varphi(x)}{T_{TOX}} - \varepsilon_{BOX}\frac{\varphi(x)}{T_{BOX}}\right) = N_{it} + \lambda_1\left(V_G - \frac{\Delta\Phi_m}{q}\right) - \lambda_2\varphi(x) \tag{3}$$

$$\lambda_1 = \frac{1}{qT_{2D}}\frac{\varepsilon_{TOX}}{T_{TOX}}, \lambda_2 = \frac{1}{qT_{2D}}(\frac{\varepsilon_{TOX}}{T_{TOX}} + \frac{\varepsilon_{BOX}}{T_{BOX}})$$

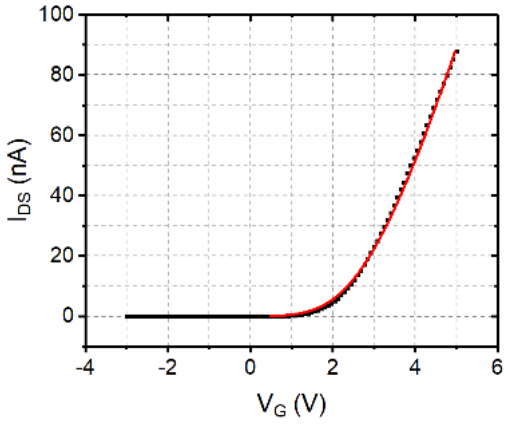
Where T_{TOX} and T_{BOX} are the thickness of the gate dielectric layer and substrate insulation layer, respectively, V_G is gate voltage, $\Delta\Phi_m = 5.54\,eV$ is the work function of metal gate. ε_{TOX} is the dielectric constant of HfO_2 gate dielectric layer, ε_{BOX} is the dielectric constant of Al_2O_3 substrate, T_{2D} is the thickness of monolayer MoS_2. According to the drift-diffusion law, carrier transport can be described as follows:

$$I(x) = qWT_{2D}[n_{2D}(x) + N_{it}(x)]\mu(x)\frac{dV_F(x)}{dx} \tag{4}$$

where W is the channel width and $\mu(x)$ is the carrier mobility. Integrating Eq. (4) along the channel from drain to source gives the current:

$$I_{d\to s} = \frac{qWT_{2D}\mu_{eff}}{L}\left(\left(N_{it} + \lambda_1(V_G - \Delta\Phi_m/q) + \frac{k_BT}{q}\lambda_2\right)(\varphi_D' - \varphi_s') - \lambda_2\frac{(\varphi_D'^{\,2} - \varphi_s'^{\,2})}{2}\right) \tag{5}$$

where μ_{eff} is the effective carrier mobility in the channel and L is the channel length. The I-V curve is in the form of a transcendental equation with fixed source-drain voltage. The equation can be converted into an explicit function of gate voltage with respect to the current.



Supplementary Fig. 4. Theoretical transfer curve (red) and measured data (black). The dimensions of the device are W = 90 um, L = 20 um, T_{TOX} = 20 nm, and T_{2D} = 0.8 nm.

3. Level-62 SPICE model

VTO	$1.24V$	$I0$	$6\ A/m$
AT	$3e^{-8}\ m$	BLK	0.001
BT	$1.9e^{-6}\ m$	DD	$1.4e^{-7}\ m$
VST	$2\ V$	DG	$2e^{-7}\ m$
VSI	$2\ V$	$I00$	$150\ A/m$
TOX	$3e^{-8}\ m$	EB	$0.68\ eV$
$EPSI$	18	MUO	$1.7\ cm^2/V \cdot s$
$ASAT$	1	$MU1$	$0.007\ cm^2/V \cdot s$
MUS	$1\ cm^2/V \cdot s$	MMU	1.3
ETA	4.5	$THETA$	$0\ m/V$
RD	$0\ \Omega$	$DELTA$	2
RS	$90000\ \Omega$	ME	2.5

Supplementary Table 1. The parameters of the MoS_2 level-62 SPICE model.

1. $VT0$

$VT0$ is defined as the turn-on voltage of the transistor at zero bias. For thin-film transistors, the effective turn-on voltage fluctuates around $VT0$. The value of $VT0$ on the transfer characteristic curve refers to the voltage applied to the gate electrode required to induce the channel region into a conductive state.

2. AT and BT

The parameters AT and BT are both used to describe the influence of drain-induced barrier lowering (DIBL) on the turn-on voltage. The initial value of AT is 3×10^{-8} m/V, and the initial value of BT is 1.9×10^{-6} m · V. When the channel length L_{eff} is sufficiently large, the magnitude of V_{teff} is approximately a constant equal to $VT0$.

3. VST and VSI

The parameters VST and VSI are used to adjust the effect of V_{gs} on the threshold voltage. When VST increases, the transfer effective turn-on voltage decreases relatively, and the subthreshold area in the characteristic curve shifts to the right, even if the threshold voltage decreases. The initial values of VST and VSI are 2 V.

4. I_0

I_0 is defined as the leakage scaling constant. The leakage current is more sensitive to changes in I_0 when V_{DS} is larger.

5. *BLK*

BLK represents the degree of contribution of V_{DS} to I_{leak}. Its physical meaning is the potential barrier formed by the leakage current, which is used to adjust the interval between the transfer characteristics of the transistor.

6. *DD*

DD reflects that the drain terminal voltage V_{DS} affects the electric field distribution at the drain terminal, thereby affecting the carrier emissivity. The initial

value is $1.4\text{x}10^{-7}$ m.

7. *DG*

DG also quantifies how the drain terminal voltage V_{GS} affects the longitudinal electric field distribution near the gate, and thus the carrier emissivity. The initial value is $DG = 2\times10^{-7}$ m.

8. MUS

MUS characterizes the mobility in the subthreshold region in units of $\text{cm}^2/\text{V}\cdot\text{s}$.

9. ETA

ETA is a model parameter used to adjust the slope of the transfer current in the subthreshold region. *ETA* is a dimensionless number with an initial value of 7. The expression of the leakage current in the cut-off level 62 TFT model is:

$$I_{leak} = I_0 \cdot W_{eff}\left[\exp\left(\frac{q\cdot BLK\cdot V_{DS}}{k\cdot T}\right) - 1\right]\cdot[X_{TFE}(F) + X_{TE}] + I_{\text{leakge}} \quad (6)$$

Linear area:

$$I = \mu_{FET}\cdot C_{ox}\cdot\frac{W_{eff}}{L_{eff}}\cdot\left(V_{GTE}\cdot V_{DS} - \frac{V_{DS}^2}{2ASAT}\right) \quad for\ V_{DS} < ASAT\cdot V_{GTE} \quad (7)$$

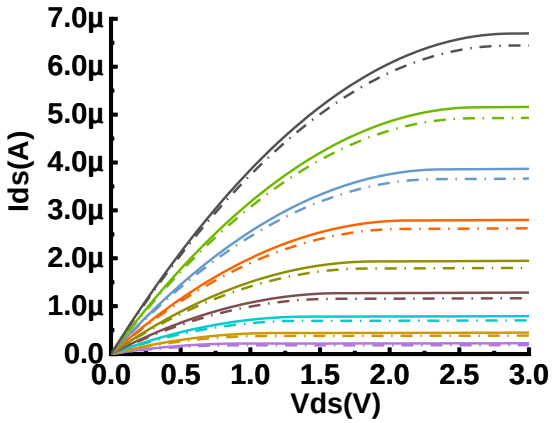
Saturation zone:

$$I = \frac{\mu_{FET}\cdot C_{ox}\cdot W_{eff}\cdot V_{GTE}^2\cdot ASAT}{2L_{eff}}\ for\ V_{DS} > ASAT\cdot V_{GTE} \quad (8)$$

where, $$V_{GTE} = V_{sth}\cdot\left[1 + \frac{V_{GT}}{2\cdot V_{sth}} + \sqrt{DELTA^2 + \left(\frac{V_{GT}}{2\cdot V_{sth}} - 1\right)^2}\right] \quad (9)$$

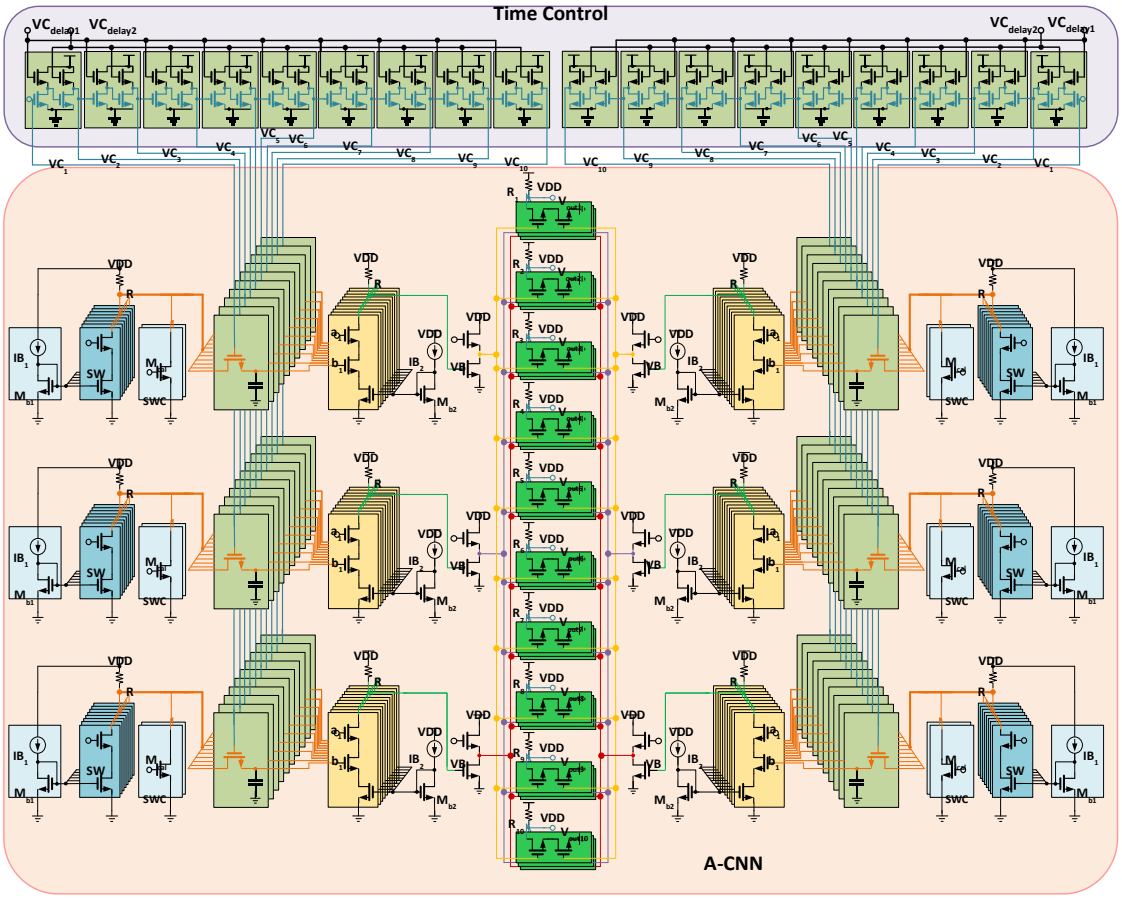
In Eqs. (7) and (8), the parameter $ASAT$ is introduced to adjust the relationship between the drain saturation voltage V_{DSAT} and the gate voltage V_{GTE}. The initial value is $ASAT = 1$, $V_{DSAT} = ASAT\cdot V_{GTE}$. μ_{FET} is the charge mobility in the channel,

C_{ox} is the equivalent gate oxide layer capacitance per unit area, L_{eff} is the effective channel length, and W_{eff} is the effective channel width. Finally, μ_{FET} in Eqs. (7-8) is the field effect mobility of the transistor when the device is turned on.



Supplementary Fig. 5. Measurements and 62 level SPICE model results.

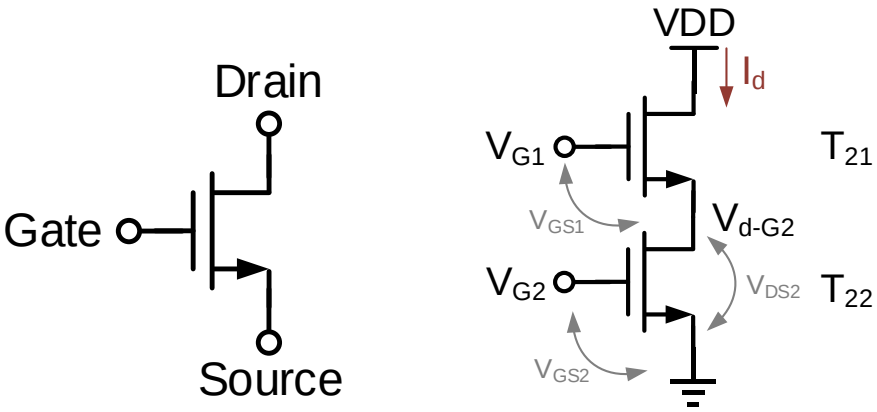
4. ANN IC structure and MAC operation analysis



Supplementary Fig. 6 Schematic of the ANN integrated circuit with various functional modules.

The above shows a schematic diagram of our MoS_2 ANN IC which consists of convolution calculation, memory, activation function, integrated weight update, and time control circuits.

Multiplication operation:



Supplementary Fig. 7 Schematic structure of a normal MoS_2 FET (left) and a dual gate transistor for multiplication operation (right).

A normal MoS_2 FET is a three-port device. In this paper the presented transistor (T_2) capable of multiplication operation is a dual-gate structured device, and the model can be simplified as transistor T_{21} and transistor T_{22} in series. V_{GS2} is the voltage difference between the gate and source of transistor T_{22}, and V_{DS2} is the voltage difference between the drain and source of transistor T_{22}. When $V_{GS2} > V_{th-G2}$ and $V_{DS2} < V_{GS2} - V_{th-G2}$, transistor T_{22} operates in the linear region. The current I_d passed by transistor T_{22} is expressed as follows:

$$I_d = \frac{\mu_n \cdot C_{ox} \cdot W}{2L} \cdot [2(V_{GS2} - V_{th-G2})V_{DS2} - V_{DS2}^2] \quad (10)$$

where μ_n is the free electron mobility, C_{ox} is the gate capacitance per unit area, and W and L are the channel width and length, respectively. When V_{DS2} in (1) is very small, the quadratic term V_{DS2}^2 can be ignored, and Eq. (1) can be simplified as follows:

$$I_d = \frac{\mu_n \cdot C_{ox} \cdot W}{L} \cdot (V_{GS2} - V_{th-G2})V_{DS2} \quad (11)$$

As shown in supplementary Fig. 7, the gate of transistor T_{21} is connected to the input signal V_{G1}, and transistor T_{21} operates in the saturation region. According to

Kirchoff's voltage law, the relationship between V_{G1} and V_{DS2} is

$$V_{DS2} = V_{G1} - V_{th-G1} \tag{12}$$

and

$$V_{G2} = V_{GS2} \tag{13}$$

Substituting (12) and (13) into (11) gives:

$$I_d = \frac{\mu_n \cdot C_{ox} \cdot W}{L} \cdot (V_{G2} - V_{th-G2})(V_{G1} - V_{th-G1}) \tag{14}$$

where $\frac{\mu_n \cdot C_{ox} \cdot W}{L}$ is a constant, which can be recorded as β, so:

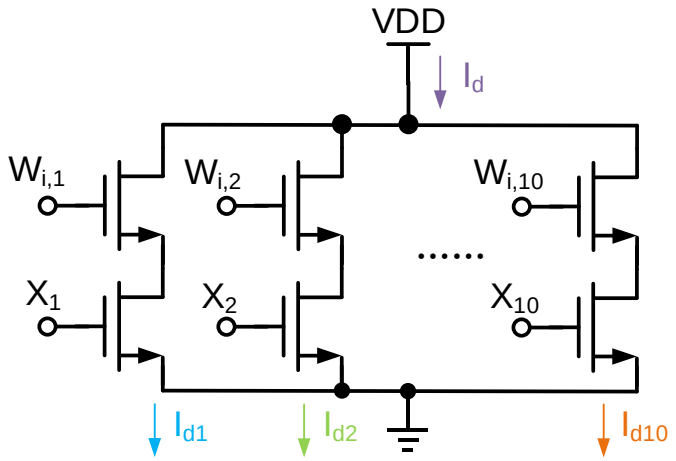
$$I_d = \beta (V_{G2} - V_{th-G2})(V_{G1} - V_{th-G1}) \tag{15}$$

where β is a constant and the current I_d has a linear correlation with the product of the input signal $(V_{G1} - V_{th-G1})$ and weight signal $(V_{G2} - V_{th-G2})$, ***giving a completed multiplication operation.***

In Eq. (15), $(V_{G2} - V_{th-G2})$ can be used as a weight w_{ij}, and (V_{G2}-V_{th-G1}) can be used as the input signal x_j, therefore:

$$I_d = \beta \cdot x_j \cdot w_{ij} \tag{16}$$

Addition operation:



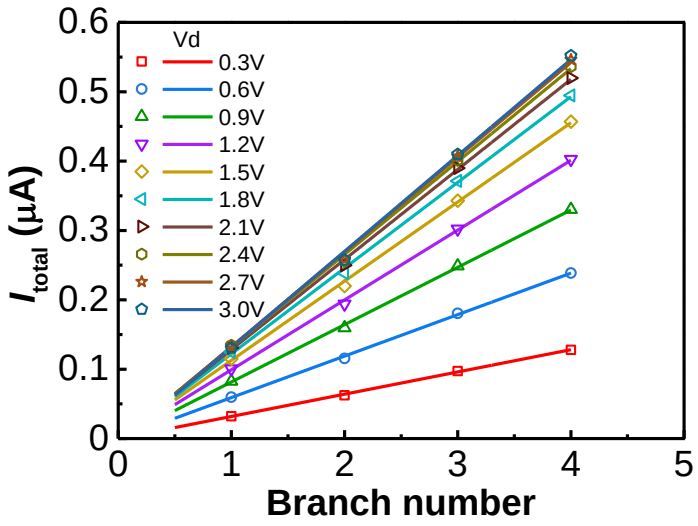
Supplementary Fig. 8 Schematic showing a circuit for an addition operation.

The current I_d is given by

$$I_d = I_{d1} + I_{d2} + \cdots + I_{d10} \tag{17}$$

$$I_d = \beta \cdot x_1 \cdot w_{i1} + \beta \cdot x_1 \cdot w_{i1} + \cdots + \beta \cdot x_{10} \cdot w_{i10} \tag{18}$$

$$I_d = \beta \cdot (x_1 \cdot w_{i1} + x_1 \cdot w_{i1} + \cdots + x_{10} \cdot w_{i10}) \tag{19}$$

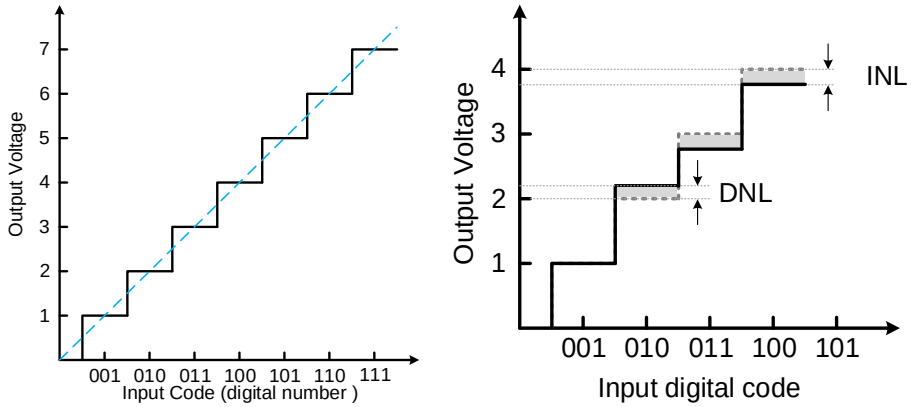


Supplementary Fig. 9 Total current as a function of the number of multiplier branches for different drain voltage values.

As shown in supplementary Fig. 9, when the transistor T2 has a different drain voltage, the current summation of all branches is also different with an increasing slope. Changing the drain voltage can suit different application scenarios.

5. Calibration: Non-ideal weight correction

A digital-to-analog converter is one of the key modules used for weight updates. Its function is to convert a digital input to an analog output. This conversion is a linear conversion. When a digital code is input, the output is proportional to its analog value. As shown below, the performance of a digital-to-analog converter is mainly characterized by its differential nonlinearity (DNL) and integral nonlinearity (INL).



Supplementary Fig. 10 Working principle of DACs.

For an N-bit DAC, the digital input range is 0 to 2^{N-1}. When the digital input changes by "1", the value of the analog output change is called the minimum quantized analog increment. The value is characterized by the Least Significant Bit (LSB). Assuming its maximum analog value is *X*, its ideal minimum quantized analog increment is:

$$1\ \mathrm{LSB} = \frac{X}{2^{N}-1} \tag{20}$$

Because of the non-linearity of the DAC, when the input digital quantity changes by "1", its output value may not be equal to 1 LSB. This error can be characterized by DNL. DNL refers to the ratio of the difference between two adjacent analog output values:

$$\mathrm{DNL}_K = \frac{X_K - X_{K-1} - 1\ \mathrm{LSB}}{1\ \mathrm{LSB}} \tag{21}$$

where X_K and $X_{K\text{-}1}$ are the analog outputs corresponding to input digital quantities K and K-1, respectively. Differential nonlinearity shows the uniformity of the analog output from the digital-to-analog converter when the digital input changes. If the adjacent input digital codes changes and its corresponding analog output changes by 1 LSB, then the output of the digital-to-analog converter (DAC) is ideally uniform. However, due to processing fluctuations and transistor uniformity, the actual DNL is usually larger than 1 LSB. The larger the DNL is, the more nonlinearity the DAC performs. Higher quality MoS_2 film, better fabrication recipes and more matched design methods are required to reduce DNL.

Due to the unavoidable nonlinearity of digital-to-analog conversion, there is a deviation between the ideal analog output and actual analog output. This deviation is characterized in terms of INL. For example, when the digital quantity K is input, the actual output is X_K, and the ideal output is $\mathrm{X}_K{}' = \frac{X}{2^N - 1} \times K$. Therefore,

$$INL_K = \frac{\mathrm{X}_K - \mathrm{X}_K{}'}{\mathrm{X}_K{}'} \tag{22}$$

The integral linear error is closely related to the differential linear error as follows:

$$INL_K = \frac{\sum_{i=1}^{K} DNL_i}{K} \tag{23}$$

The error of DNL and INL mainly affects the weight update accuracy, which directly affects the convergence speed. Although the error caused by the fluctuations in the process cannot be completely overcome with a symmetric layout, the error in INL and DNL can be further overcome by using a calibrated current source. For example, when the actual output current is larger than the theoretical value, the output current

can be made closer to the theoretical value by reducing the current from the current source. As shown in Fig. 5h in the maintext, the step error is 1 LSB when the calibration circuits are turned on, and the calibrated INL and DNL values shows that the DAC has an 8-bit accuracy.

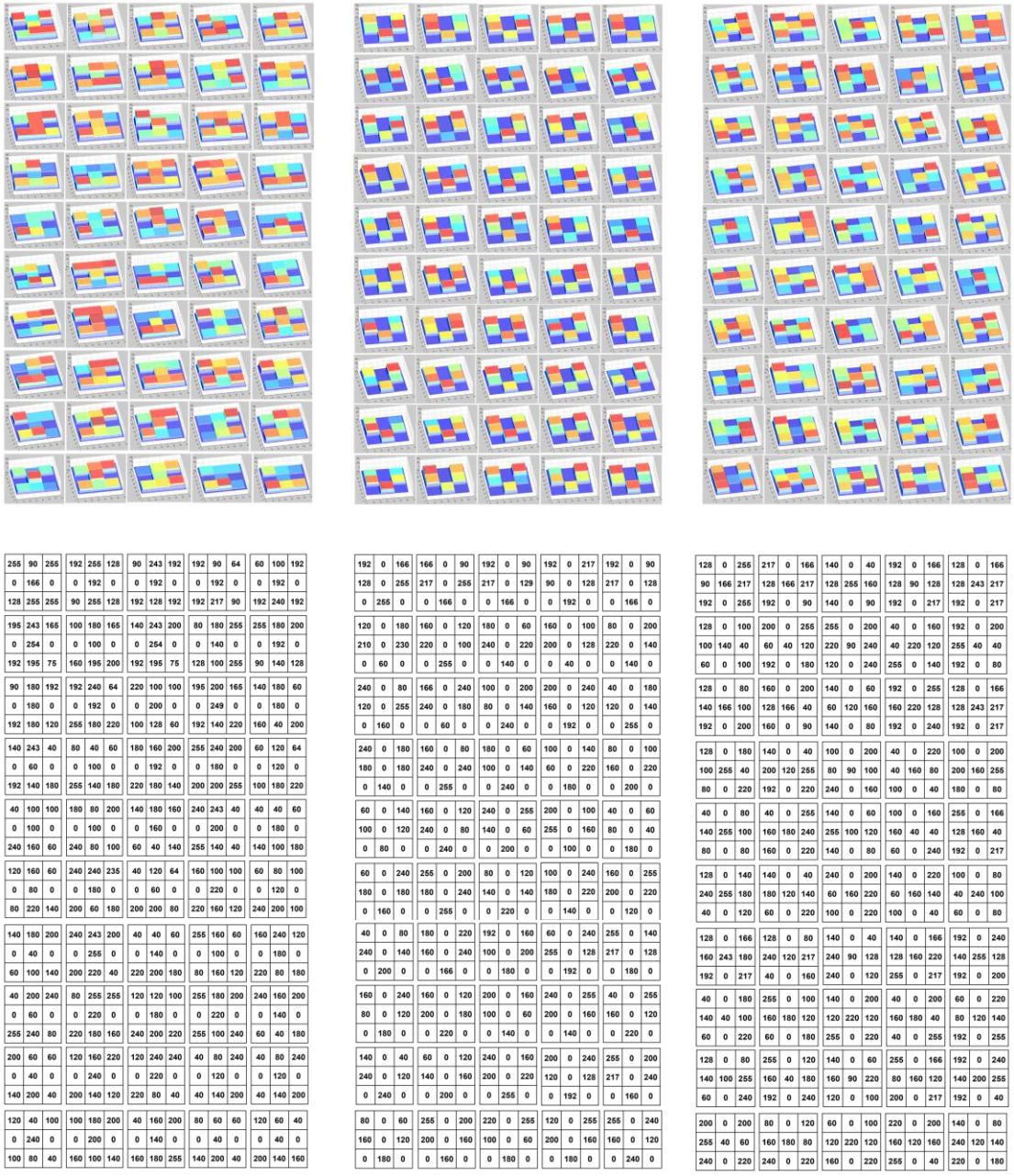
6. Artificial neural networks and deep learning algorithm

Convolutional neural networks can be divided into two parts. One part is the convolution operation that is the basis for deep learning, and the other part is an activation function part, which provides a judgment of the network update. The use of back propagation algorithms provides weight convergence and adjustment. The input of the convolutional layer is composed of pixels of the image and the output of the convolutional layer represents the extracted feature of the input image. The purpose of the fully-connected layer is to use these features to classify the input image.

The output from the neural network layer uses the activation function to determine the final output. The input from the activation function is the input to the previous network layer, which is a vector of values between 0 and 1 of any value greater and converts them into a vector. The commonly-used activation functions are the sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU) function. In this study, continuously differentiable functions are used for activation. The ReLU function has several important disadvantages. For example, if the learning rate is too high, neurons are not activated during the training process, resulting in a neural network with greatly reduced efficiency. Therefore, we present an activation function that can overcome this drawback.

For application of such convolutional neural network, a normalization of input data is necessary. The 8-bit pressure sensing data ranges from 0 to 255. Since the MoS_2

FET provides a high linearity for small bias signals, the magnitude of the input voltage determines the gray scale range. Therefore, a linear approximation is applied to the input image. After such pre-processing, the image is now transformed to digital data and can be used in a convolution operation. Each image can be viewed as a matrix of pixel values, as shown in the supplementary Fig. 12,



Supplementary Fig. 12 The data set used for the ANN training.

The use of deep learning algorithms to adjust weights is the key of identification and classification. As an example, if the letter Z was used as an input, the target probability for the class Z is 1, while the target probability for the other three classes is 0, namely:

- Image = Z
- target vector = [0, 0, 1]

The process for training the convolutional neural network can be summarized as follows:

Step 1: The training library is produced by a simple program, and we initialize all weights with random values. The data of the corresponding matrix are shown in supplementary Fig. 12. Instead of simply using 0 or 1 to represent the pixel, our image also 256 gray levels for each pixel to improve the accuracy of the neural network learning process.

Step 2: The neural network receives a training image as an input, and the probability the image corresponds to each class is determined using forward propagation (convolution, activation in the and hidden layer, and forward propagation to the fully connected layer). The target matrix corresponds to vectors [0, 0, 1] for Z, [0, 1, 0] for N, and [1, 0, 0] for V.

Step 3: Calculate the total error for the output layer (calculate the sum of the 3 categories)

$$\text{Total Error} = \sum \tfrac{1}{2} (\text{target value} - \text{output value})^2$$

Step 4: Use BP to calculate the gradient according to the output total error of the network, and use the gradient descent algorithm to update the values/weights of all

filters and parameter values to minimize the output error. In terms of a unit weight w_{ij}, the required correction for each step given by the steepest descent method is

$$\Delta w_{\mathrm{ij}} = -\eta \frac{\partial E}{\partial w_{ij}} = \eta \sum_s [T_i^s - O_i^s] \varphi'(h_i^s) H_j^s = \eta \sum_s \delta_i^s H_j^s \tag{24}$$

where

$$\delta_i^s = \varphi'(h_i^s)[T_i^s - O_i^s] \tag{25}$$

The weight from an input cell to a hidden cell w_{jk} is

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = \eta \sum_{s,i} [T_i^s - O_i^s] \varphi'(h_i^s) w_{ij} \varphi'\left(h_j^s\right) I_j^s = \eta \sum_{s,i} \delta_i^s w_{ij} \varphi'\left(h_j^s\right) I_k^s = \eta \sum_s \bar{\delta}_j^s I_k^s \tag{26}$$

where

$$\delta_j^s = \varphi'\left(h_j^s\right) \sum_i w_{ij} \delta_i^s \tag{27}$$

This iterative algorithm aims to ensure $\{w_{ij}, w_{jk}\}$ provides high accuracy. The value of η is the learning efficiency.

Step 5: Repeat steps 1 to 4 for all the images in the training dataset.