Transformer-Based Nuclide Identification Method for γ Spectrum and Model Construction

Authors: Fei Li, Chu-Yang Luo, Ying-Zi Wen, Sheng Lv, Feng Cheng, Guo-Qiang Zeng, Jian-Feng Jiang, Bing-Hai Li, Feng Cheng, Guo-Qiang Zeng

Date: 2024-07-30T00:00:00+00:00

Abstract

In current neural network algorithms for nuclide identification in high-background, poor-resolution detectors, traditional network paradigms including back-propagation networks, convolutional neural networks, recurrent neural networks, etc. have been limited in research on γ spectrum analysis because of their inherent mathematical mechanisms. It is difficult to make progress in terms of training data requirements and prediction accuracy. In contrast to traditional network paradigms, network models based on the transformer structure have the characteristics of parallel computing, position encoding, and deep stacking, which have enabled good performance in natural language processing tasks in recent years. Therefore, in this paper, a transformer-based neural network (TBNN) model is proposed to achieve nuclide identification for the first time. First, the Geant4 program was used to generate the basic single-nuclide energy spectrum through Monte Carlo simulations. A multi-nuclide energy spectrum database was established for neural network training using random matrices of γ-ray energy, activity, and noise. Based on the encoder-decoder structure, a network topology based on the transformer was built, transforming the 1024-channel energy spectrum data into a 32×32 energy spectrum sequence as the model input. Through experiments and adjustments of model parameters, including the learning rate of the TBNN model, number of attention heads, and number of network stacking layers, the overall recognition rate reached 98.7%. Additionally, this database was used for training AI models such as back-propagation networks, convolutional neural networks, residual networks, and long short-term memory neural networks, with overall recognition rates of 92.8%, 95.3%, 96.3%, and 96.6%, respectively. This indicates that the TBNN model exhibited better nuclide identification among these AI models, providing an important reference and theoretical basis for the practical application of transformers in the qualitative and quantitative analysis of the γ spectrum.

Full Text

Preamble

A Transformer-Based Nuclide Identification Method for γ-Ray Spectra and Model Development

Fei Li,[1,2,3] Chu-Yang Luo,[1] Ying-Zi Wen,[1] Sheng Lv,[1] Feng Cheng,[1,2,†] Guo-Qiang Zeng,[1,2,‡] Jian-Feng Jiang,[1] and Bing-Hai Li[4,5]

[1]College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu 610000, China
[2]Applied Nuclear Technology in Geosciences Key Laboratory of Sichuan Province, Chengdu 610000, China
[3]Department of Physics, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain
[4]Airborne Survey and Remote Sensing Center of Nuclear Industry, Shijiazhuang 050000, China
[5]Hebei Key Laboratory of Airborne Detection and Remote Sensing Technology, Shijiazhuang 050000, China

Current neural network algorithms for nuclide identification in high-background, poor-resolution detectors have encountered limitations due to the inherent mathematical mechanisms of traditional network paradigms such as back-propagation networks, convolutional neural networks, and recurrent neural networks. These constraints have hindered progress in both training data requirements and prediction accuracy for γ-ray spectrum analysis.

In contrast to conventional network paradigms, transformer-based models feature parallel computing, positional encoding, and deep stacking capabilities, which have demonstrated exceptional performance in natural language processing tasks in recent years. Therefore, this paper proposes, for the first time, a transformer-based neural network (TBNN) model for nuclide identification. First, the Geant4 program was employed to generate basic single-nuclide energy spectra through Monte Carlo simulations. A multi-nuclide energy spectrum database was then constructed for neural network training using random matrices of γ-ray energy, activity, and noise. Based on the encoder-decoder architecture, a transformer-based network topology was developed, transforming the 1024-channel energy spectrum data into a 32×32 energy spectrum sequence as model input. Through systematic experimentation and parameter tuning—including the TBNN model's learning rate, number of attention heads, and network stacking layers—the overall recognition rate reached 98.7%. When the same database was used to train AI models such as back-propagation networks, convolutional neural networks, residual networks, and long short-term memory neural networks, the overall recognition rates were 92.8%, 95.3%, 96.3%, and 96.6%, respectively. These results demonstrate that the TBNN model achieved superior nuclide identification performance among these AI models, providing an important reference and theoretical foundation for the practical application

of transformers in qualitative and quantitative γ-ray spectrum analysis.

Keywords: Nuclide identification; Neural network; Transformer

Introduction

Multi-nuclide identification is a critically important radioactive material detection technology in medicine, national defense, and social stability [**?**, **?**]. For γ-ray spectra acquired under complex backgrounds with low count rates and poor signal-to-noise ratios, multi-nuclide identification presents significant challenges [**?**]. Traditional machine learning-based methods are suitable only for scenarios with few nuclide types (4) and small datasets [**?**], suffering from low recognition rates and inability to achieve rapid identification. Conventional shallow neural network structures tend to overfit and exhibit poor generalization during training, whereas deep network structures can alleviate these problems [**?**] but involve numerous hyperparameters and are difficult to train.

In the early days, due to limited computational capacity, nuclide identification primarily relied on traditional peak-searching methods based on full-energy peaks, such as maximum value, symmetric zero-area conversion, and derivative methods [**?**, **?**]. However, these traditional approaches are not applicable to situations with low count rates or low peak-to-background ratios [**?**].

Machine learning classification algorithms—including Bayesian, decision tree, and support vector machine algorithms, along with their variants—have been applied to nuclide identification. Compared with traditional methods, these approaches offer advantages in recognition speed and accuracy under low-count-rate conditions [**?**]. However, machine learning algorithms require researchers to manually extract data features [**?**, **?**], such as peak positions and boundaries of characteristic peaks, spectrum background, and noise levels, before making decision inferences based on these extracted features and performing classification statistics.

The accuracy of nuclide identification depends on continuous improvement of feature extraction algorithms, yet increasingly complex feature extraction steps raise the difficulty of identification [**?**]. Multi-layer perceptrons (MLP) in machine learning have been applied in nuclide identification and signal processing [**?**, **?**], demonstrating good approximation and generalization capabilities. While overall prediction accuracy for nuclide identification tasks can exceed 90%, MLPs are prone to falling into local optima. Combining feature extraction algorithms with MLP can effectively improve the system and increase nuclide identification accuracy by 5%–9% [**?**]. For example, Yicong et al. used an improved particle swarm algorithm to optimize the threshold and weight values of a back-propagation network (BP) [**?**]. Nevertheless, these machine-learning-based nuclide identification methods exhibit a sharp decline in recognition rate when dealing with more than five radioactive nuclides [**?**], and their dependence on manual feature extraction increases the difficulty of identification.

In recent years, with improvements in hardware performance such as graphics processing units (GPUs) and tensor processing units (TPUs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and extended deep learning models based on CNNs or RNNs have begun to demonstrate advantages. As a branch of machine learning, deep learning has been widely applied in image recognition, target detection, and semantic segmentation due to its simplified tensor operations, good scalability, and universality [**?**, **?**]. Deep learning models typically consist of multiple layers, each fully connected to those below (from which they receive input) and above (which they influence). The entire model and its constituent layers share this structure, using raw inputs (features), generating outputs (predictions), and possessing parameters (combined from all constituent layers). Similarly, each individual layer receives inputs from the previous layer, generates outputs for the subsequent layer, and possesses tunable parameters updated according to signals flowing backward from subsequent layers.

To date, researchers have explored the potential applications of deep learning models in nuclide identification, with research focusing on three aspects: dataset construction, input data preprocessing, and network model improvement [**?**]. Increasing network depth improves nuclide identification performance to some extent, as demonstrated by using ResNet-50 for full-spectrum all-nuclide identification [**?**]. However, the large number of hyperparameters (10^7) makes model training time-consuming and difficult to optimize, with high hardware requirements. The performance of current network models in nuclide identification—including BP, CNNs, residual networks (ResNet), and long short-term memory neural networks (LSTMs)—is limited by their inherent mathematical mechanisms, making breakthroughs difficult in terms of training data requirements and prediction accuracy. Therefore, applying new network model paradigms to nuclide identification tasks represents a potential avenue for improvement.

The transformer is a neural network architecture distinct from traditional models such as RNNs or CNNs. It employs techniques such as residual connections and layer normalization to accelerate parallel computing [**?**] and is commonly used for natural language processing (NLP) tasks like language translation, language modeling, and sentiment analysis. For instance, the ChatGPT model, which received significant attention in late 2022, was improved based on the transformer architecture. The key innovation of the transformer is its self-attention mechanism, which enables networks to focus on different parts of the input sequence at different computational stages without relying on a fixed-length context window. This allows networks to dynamically attend to different parts of the input sequence based on their relevance to the task at hand [**?**].

Although originally introduced for NLP tasks, researchers have also explored the transformer's potential in image recognition. One major advantage of using transformers for image recognition is their ability to capture long-range correlations in input images. Through the self-attention mechanism, transformers can learn to extract features useful for image recognition tasks [**?**, **?**]. Dosovit-

skiy et al. proposed a Vision Transformer (ViT) method combining CNNs with transformers for image recognition tasks, while Si et al. developed an Inception Transformer (iFormers) architecture to address the transformer's deficiency in capturing high-frequency local information [**?**, **?**]. Both ViT and iFormers have demonstrated excellent results in various image recognition tasks, including classification, segmentation, and object detection. Since deep learning-based nuclide identification methods use full-spectrum data as input, which can be treated as either one-dimensional sequence data or transformed into two-dimensional image input, the transformer architecture is theoretically applicable to nuclide identification tasks.

This paper proposes, for the first time, the use of a transformer model to replace traditional network paradigms for nuclide identification and explores the potential application of transformers in this domain. This study verifies the method's scalability, stable gradient propagation capability, high accuracy in spectrum nuclide identification, and robustness through training and testing with simulated spectrum data. Furthermore, we provide a comparison with traditional BP, CNN, and RNN models in terms of full-spectrum recognition rate and convergence speed.

II. Algorithm and Model

A. Construction of Training Set Through Monte Carlo Simulation

According to the IAEA-2006 standard industrial nuclide library [**?**], the radioactive nuclides selected for identification in this study were ^{241}Am, ^{192}Ir, ^{226}Ra, ^{133}Ba, ^{60}Co, ^{57}Co, ^{137}Cs, and ^{152}Eu. Monte Carlo (MC)-based methods are commonly used in nuclear science and detection to simulate radioactive nuclide spectra. Geant4 is a software package developed by the European Organization for Nuclear Research (CERN) using the C++ platform, widely applied in nuclear physics, radiation protection, and detection [**?**]. This study utilized Geant4 to simulate the NaI detector response in nuclide spectrum experiments, employing a standard cylindrical NaI detector with dimensions 5×5 cm.

The radiation source was configured as a point source located 5–15 cm (randomly selected) directly in front of the detector, with 1.0×10^8 emitted particles. To account for varying energy resolutions of detectors in real scenarios, this study adopted Gaussian broadening according to the following formula:

$$E_{new} = \text{Gauss}(x_{Random}; E_i, \delta_{FWHM})$$

$$\delta_{FWHM} = a + bE_i + cE_i^2$$

where x_{Random} is a random number, E_i is the initial energy of the i-th channel, and a, b, c are the detector broadening coefficients [**?**]. After obtaining simulated radioactive nuclide spectrum data through MC simulation, we acquired

initial 80 spectrum datasets (ten spectra per nuclide), each containing 1024 channels (0–3 MeV). Considering the activity levels of radioactive nuclides and environmental noise impacts in actual measurements, this study expanded the nuclide spectrum dataset by adding random noise and amplitude transformations, using random combinations to obtain mixed spectra from multiple source nuclides. The calculation formula is as follows:

$$\mathrm{Spec}_{new} = \mathrm{Add\_noise}\{\mathrm{Spec}_{^{241}\mathrm{Am}} \times \mathrm{random}(0.1{-}10)^{\mathrm{random}[0,1]} + \mathrm{Spec}_{^{192}\mathrm{Ir}} \times \mathrm{random}(0.1{-}10)^{\mathrm{random}[0,1]} + \cdots + \mathrm{Spec}_{1}$$

where Add_{noise} refers to adding Gaussian white noise with a signal-to-noise ratio (SNR) between 25 and 35, random(0.1–10) is a random number from 0.1 to 10, random[0,1] refers to a random value of 0 or 1, and Spec is the corresponding nuclide spectrum. The spectrum dataset for this study was constructed in this manner with a database volume of 1.0×10^4.

B. Construction of Neural Network Models

1. Transformer-Based Neural Network (TBNN) Nuclide Identification Model The attention mechanism can be considered a simulation of human visual processing. Since visual information processing consumes substantial brain resources, the brain does not process all information at the highest granularity but instead focuses on regions of interest to allocate resources efficiently [**?**]. The multi-head attention mechanism derives from the self-attention mechanism, which constitutes the core of the transformer.

As shown in Fig. 1 [Figure 1: see original paper], the self-attention mechanism is a top-down approach for calculating similarity and feature associations between different data [**?**]. The self-attention mechanism can be represented by the following formulas:

$$Q, K, V = sW^Q, sW^K, sW^V$$

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

where $s \in \mathbb{R}^{l \times h}$ represents an input sequence with sequence length l and feature dimension h, and W^Q, W^K, W^V are fully connected layers of size $h \times k$ that map each input s to q, k, v vectors of size $1 \times k$, respectively, combining them into Q, K, V matrices of size $l \times k$. By calculating corresponding attention scores α using the dot product of q and each k vector, then summing the corresponding v vectors (weighted by each normalized attention score), we obtain a new output corresponding to this q. In matrix representation, the attention score matrix of size $l \times l$ is $\mathrm{softmax}(QK^T/\sqrt{k})$.

The multi-head attention mechanism calculates n_{head} attention vectors by dimensionally increasing through a fully connected layer to $n_{head} \times$ dimension, then uses a fully connected layer to fuse features calculated through multi-head attention. The multi-head attention mechanism can be represented by the following formulas:

$$Q_{nhead}, K_{nhead}, V_{nhead} = QW^{Q}_{nhead}, KW^{K}_{nhead}, VW^{V}_{nhead}$$

$$\text{Multihead} = \text{Attention}(Q_{nhead}, K_{nhead}, V_{nhead})$$

$$\text{MultiHead}(Q, K, V) = \text{MultiHead} \cdot W^{0}$$

where W^{Q}_{nhead}, W^{V}_{nhead}, W^{K}_{nhead} correspond to the multi-head dimensional matrices of Q, K, V respectively, and W^{0} is a dimensional reduction matrix used for fusing multi-head attention features.

The transformer is a model based on the multi-head attention mechanism, where each sub-layer uses residual connections followed by layer normalization. The transformer-based neural network (TBNN) model designed in this study is shown in Fig. 2 [Figure 2: see original paper]. During the training stage, the input "Targets" in the decoder portion represents the predicted targets corresponding to the dataset and is used to enforce teacher forcing on the network. During the prediction stage, the decoder performs cyclic prediction until 32 sequence predictions are completed, finally outputting the result through the "FN+Flatten" layer.

The spectrum output from the NaI detector consists of 1024 channels. The model first transforms the 1×1024 spectrum into a 32×32 sequence, eliminating the need for the embedding layer used in the original transformer model to encode the 1024 data points. Positional encoding adds absolute or relative positional information to the input sequence. Fixed positional encoding based on sine and cosine functions was employed, with the following calculation formulas:

$$\text{PE}_{m,2n} = \sin\left(\frac{m}{10000^{2n/32}}\right)$$

$$\text{PE}_{m,2n+1} = \cos\left(\frac{m}{10000^{2n/32}}\right)$$

where $\text{PE}_{m,2n}$ and $\text{PE}_{m,2n+1}$ represent the positional encoding calculation results for indices $(m, 2n)$ and $(m, 2n+1)$ in the sequence, respectively. Adding them to the sequence before encoding yields the output of the positional encoding layer.

The output of the positional encoding layer enters the encoder-decoder structure, after which the sequence is flattened and connected to a fully connected layer, finally outputting a nuclide identification result with dimension 1×8.

2. Other Neural Network Models for Comparison To evaluate the performance of the newly constructed model, this study selected four neural network models widely used in nuclide identification for comparison: BP, CNN, ResNet, and LSTM [**?**, **?**].

A CNN is essentially a multi-layer perceptron composed of convolutional layers, pooling layers, input layers, fully connected layers, and output layers [**?**]. The convolutional and pooling layers in the hidden layer form the core of the CNN for feature extraction. The convolution operation in the convolutional layer achieves local connections; its output data are calculated for each neuron using the same convolution kernel (shared weights) and then added with the same bias (shared bias). The pooling layer contains a structure similar to a convolutional kernel for performing pooling operations on data but has no learnable parameters; it only takes the maximum or average value from the target region. The pooling operation utilizes the principle of local image correlation to sample data, increasing the CNN's robustness to small positional changes while retaining useful information. These two main layers effectively reduce the number of network parameters and alleviate model overfitting. Assuming the input is N, the convolution kernel is L, and the feature mapping is M, the corresponding two-dimensional convolution operation expression is as follows:

$$S(n,m) = (N \cdot L)(n,m) = \sum_i \sum_j N(i,j)L(n-i, m-j)$$

where I is the width of the convolution kernel and J is the height of the convolution kernel.

ResNet is an improved CNN model built from residual blocks that form a residual network [**?**]. Residual connections allow the network to reach more ancestors during back-propagation, thereby alleviating network degradation. ResNet performs well when training deep networks [**?**], enabling effective training of deep neural networks. The main formula for the residual network is:

$$F(x) = g(x) + x$$

where $F(x)$ is the final output of ResNet, $g(x)$ is the output of the two convolutions, and x is the sample data.

LSTM was proposed by Hochreiter and Schmidhuber in 1997 to address problems faced by traditional RNN models, which have difficulty learning long-term dependencies and are prone to gradient vanishing and explosion [**?**]. The LSTM

model performs selective addition and deletion during data input, relying primarily on three gate structures: forget, update, and output gates [**?**]. The forget gate determines which data to discard or retain through the sigmoid layer. The update gate then screens data content and selects updates to the state, primarily determined through tanh and sigmoid layers. Finally, the output gate combines current memory with long-term memory, then judges whether this result should be output through the sigmoid activation function layer, passing it to the next neuron cell. The sigmoid and tanh layers use sigmoid and tanh functions as activation functions, respectively, to enhance nonlinear relationships between neurons.

Theoretically, a three-layer BP neural network (with a concise structure and relatively few parameters) can approximate any given function with arbitrary accuracy, which is a compelling prospect [**?**]. Simultaneously, BP possesses certain generalization capabilities. However, from a mathematical perspective, BP is a local search method that is highly likely to fall into local extrema when solving for the global extremum of complex nonlinear functions, causing training failures [**?**, **?**]. CNNs excel at feature extraction for two-dimensional inputs, with structural characteristics of local connections, weight sharing, and pooling operations [**?**, **?**]. Due to the balance between local feature extraction and global feature interaction, CNNs typically outperform BP [**?**]. The primary reason for using ResNet in this study was to test whether it could achieve better performance than CNNs when increasing convolutional depth. In theory, due to residual blocks, ResNet can train deeper neural networks, such as ResNet-50, which performs well on nuclide recognition tasks [**?**]. The LSTM model exhibits outstanding performance in processing sequential inputs, mitigating the long-term dependency problem in RNNs [**?**]. Each LSTM cell contains several MLPs, and if the LSTM time span is large and the network is deep, the training phase becomes computationally intensive and time-consuming (due to lack of parallel computing).

The parallel computation of the multi-head attention mechanism in the transformer model significantly improves training and inference efficiency, enabling larger models and processing of longer sequences. However, its high computational cost, structural complexity, and numerous hyperparameters increase optimization difficulty, requiring careful adjustment of hyperparameters such as learning rate and batch size to achieve better performance [**?**, **?**]. Since energy spectrum data can be regarded as both two-dimensional images and sequences, the aforementioned models were selected to test their nuclide recognition performance.

III. Results and Discussion

Eight nuclides from the industrial nuclide library were selected for analysis, and a Geant4-simulated NaI detector was used to establish a spectrum database corresponding to these eight nuclides. To make the simulated data more realistic, the database was preprocessed using Gaussian broadening and random

noise superposition. Based on data augmentation principles, the database size was expanded to 1.0×10^4. A TBNN model was constructed according to the characteristics of the input spectrum data, and traditional network paradigms including BP, CNN, and RNN were also built for comparative experiments to analyze the recognition rate, convergence speed, and other performance aspects of the TBNN model. The overall process is illustrated in Fig. 3 [Figure 3: see original paper].

A. Data Preparation

Using the method described in Section 2.1, a nuclide library was generated containing eight industrial radioactive nuclides: ^{241}Am, ^{192}Ir, ^{226}Ra, ^{133}Ba, ^{60}Co, ^{57}Co, ^{137}Cs, and ^{152}Eu. The database volume was 1.0×10^4. Each γ-ray spectrum in the database was formed using random combinations of the eight nuclides with random activities. Some γ-ray spectra are shown in Fig. 4 [Figure 4: see original paper]. Dataset annotation in this paper adopted a $[1 \times 8]$ matrix label, where each column corresponds to ^{241}Am, ^{192}Ir, ^{226}Ra, ^{133}Ba, ^{60}Co, ^{57}Co, ^{137}Cs, and ^{152}Eu, respectively. "0" indicates the absence of the radionuclide, while "1" indicates its presence. For example, if a spectrum contains ^{192}Ir, ^{152}Eu, and ^{57}Co, the label would be "[0, 0, 1, 0, 0, 1, 0, 1]." Therefore, the spectrum recognition task in this study corresponds to a multi-label classification task, with each label annotated after MC simulation.

Traditional nuclide library comparison methods are highly complex and inaccurate in identifying nuclides during the process [**?**, **?**]. In the nuclide library selected for this study, some nuclides had similar characteristic γ-ray energies, leading to overlapping peaks. For example, the difference between the characteristic γ-ray energy of ^{57}Co at 122.0614 keV (85.60%) and ^{152}Eu at 121.782 keV (39.76%) is far less than the energy resolution of the NaI scintillator detector (7%–9% at 661 keV for ^{137}Cs). As shown in Fig. 5 [Figure 5: see original paper], although these nuclides have other characteristic peaks with smaller branching ratios, it is difficult to distinguish the two nuclides from the spectrum under high-background and low-activity conditions. The database generated by Geant4 in this study used a combination spectrum method with random activity, random noise, and random nuclide types to objectively verify the TBNN model's ability to identify overlapping peaks.

Neural network models typically require a database independent of the training set to evaluate model inference capability; therefore, the simulated nuclide library was divided into training and validation sets at a 9:1 ratio. To facilitate model training, the 1×1024 spectrum data required normalization before input. Z-score normalization was used to transform the data to follow a normal distribution, with the following formula:

$$X_i^{\text{Norm}} = \frac{X_i - \mu}{\sigma}$$

where X_i denotes the i-th channel spectrum data, μ represents the mean value of the spectrum, σ represents the standard deviation of the spectrum, and X_i^{Norm} is the normalized result for the i-th channel.

B. Training of Models

1. Settings of Models for Comparison Based on the model design principles described in Section II, the models constructed in this study were built using Python with Keras and TensorFlow 2.6.0. The experimental platform was an RTX 2060 12GB GPU. During the training phase, the model performs numerous calculations such as gradient computations and parameter updates, imposing certain requirements on GPU memory size and computing performance. Additionally, overhead from model weights, gradient propagation, optimizer parameters, input data and labels, intermediate calculations, temporary buffers, hardware, and dependency libraries also consume memory. Taking the TBNN model constructed in this study as an example, if the 1×1024 energy spectrum is treated as a sequence of length 1024 during input, performing embedding operations (mapping data to a high-dimensional space) [**?**] transforms the model input shape to 1024×200 (using 200-dimensional embedding as an example). During testing, a 12GB GPU produced an "Out Of Memory (OOM)" error, indicating that the memory required to train the model exceeded available graphics card resources. Therefore, the input sequence was set to a 32×32 shape, and the embedding process was skipped entirely. By monitoring GPU memory status, we determined that the graphics card occupied approximately 1.5 GB (the maximum among all models) during the training phase.

The TBNN model designed in this study adopts a structure with 1–8 encoder-decoder layers, with the number of attention heads set to 2^n (n = 0, 1, 2, 3). The comparison models included a BP network with four hidden layers, a CNN with two convolutional layers and one fully connected layer, a ResNet with three residual blocks, and a unidirectional LSTM model with two hidden layers, as shown in Fig. 6 [Figure 6: see original paper].

The task of identifying eight radioactive nuclides can be regarded as a multi-label classification problem; therefore, the loss function for network training was binary cross-entropy (a measure of the difference between two probability distributions commonly used in binary classification problems). All models used random initialization for weight initialization, and the adaptive moment estimation (Adam) algorithm served as the optimization function. Adam combines momentum-based gradient descent and root mean square propagation (RMSProp) algorithms, reducing the number of iterations required to reach optimal values and improving optimization capability.

2. Parameters Tuning of TBNN Model a. Learning Rate

To avoid model overfitting or gradient explosion while ensuring adequate convergence speed, we tested different learning rates. The learning rates were set

to 0.01, 0.005, 0.003, 0.001, 0.0008, and 0.0005, respectively, while the number of epochs, attention head count, and layer count were fixed at 20, 4, and 4. The test results are shown in Fig. 7 [Figure 7: see original paper].

When the learning rate was set to 0.01, the model failed to converge, indicating that the initial learning rate was too large; therefore, reducing the learning rate was appropriate to obtain the optimal initial value. Better convergence speed was achieved when the learning rate was less than 0.001. Table 1 lists the test accuracy for different learning rates within 20 training steps. Consequently, subsequent model training was performed with the learning rate set to 0.0008.

Table 1. Test accuracy for different learning rates (20 epochs)

Learning Rate	Accuracy
0.01	0.6%
0.005	15.1%
0.003	75.2%
0.001	78.3%
0.0008	77.1%

b. Number of Attention Heads and Layers

Next, we analyzed the recognition rates of TBNN models with different numbers of attention heads and transformer layers. The learning rate was set to 0.0008 and the number of epochs to 20 (not considering the slowdown in convergence due to increased model parameters). Since model parameter initialization was set to random initialization, all models were trained three times, and the best recognition rate was used for horizontal comparison. The best-performing model for each combination of attention head count and network layer count was selected, with results shown in Fig. 8 [Figure 8: see original paper]. Note that the surface in the figure was obtained through linear fitting of measured points, intuitively displaying accuracy trends with parameter changes.

From the perspective of varying network layers with fixed attention head count within 20 training cycles, models with different attention head counts all achieved over 80% accuracy, reaching their current best performance with two to four network layers. However, from the perspective of varying attention head count with fixed network layers within 20 training cycles, increasing the attention head count did not significantly affect network accuracy.

From Fig. 8, optimal accuracy is achieved when the number of attention heads is two and the number of layers is four. For models with deeper network layers and more attention heads, the surge in hyperparameters makes it difficult to achieve ideal accuracy within 20 cycles. Simultaneously, the large number of hyperparameters makes network training more time-consuming and difficult to converge. Table 2 lists the training time costs for networks with different

layer counts and attention head counts, showing positive correlation with both parameters.

Table 2. Training time (seconds) for different numbers of attention heads and network layers

Layer (hyperparameters)	Head 1	Head 2	Head 4	Head 8
1 (186,700)	120	150	210	330
2 (336,016)	180	240	360	600
4 (634,648)	300	420	660	1140
8 (1,231,912)	540	780	1260	2220

c. Number of Epochs

The aforementioned training process exceeded 20 training cycles in all cases. Models with smaller parameter quantities may reach optimal fitting ability within 20 cycles; however, deeper models may be far from convergence. A reasonable number of epochs can provide good predictive ability without overfitting. Therefore, we next set the epochs to 200 to test the convergence accuracy of different models. Based on results from Sections 1) and 2), models named "(num_{Head}, num_{Layer})" were selected, including (2,2), (2,4), (2,8), and others. Each model was trained three times, recording the step length when each model reached optimal accuracy. The final step length results are shown in Table 3 and Fig. 9 [Figure 9: see original paper]. Note that Fig. 9 serves a similar purpose to Fig. 8, with the surface obtained through linear fitting of measured points to visually display accuracy trends with parameter changes.

Table 3. Number of epochs required to train each model to achieve optimal accuracy

num_{Head} num_{Layer}	2	4	6	8
2	45	60	85	120
4	50	70	95	140
8	65	90	125	180

At 20 training cycles, models with smaller parameter values (fewer attention heads and layers) achieved higher accuracy than models with larger parameter values within a shorter time span. When the number of training cycles was increased to 200, models with larger parameter values were sufficiently trained, and their accuracy also increased to above 95%. However, continuously increasing model parameters does not continuously improve accuracy, as shown in Fig. 9. When the number of attention heads was four and the number of layers was four, the model's recognition rate reached its highest value of 98.7%. This may be because an excessive number of hyperparameters creates unnecessary

or redundant connections in network layers, affecting the model's ability to extract data features and increasing training cost. However, this does not mean that models with fewer hyperparameters consistently achieve superior results. The model with parameters (4, 4) exhibited the best nuclide recognition performance for the database established in this study and was used for comparison with other neural network models, as described in the next section.

3. Comparison with Other Neural Network Models The four models shown in Fig. 6 [Figure 6: see original paper] were trained for comparison with the proposed TBNN model. For BP, which has fewer network nodes, a large step size can easily lead to overfitting; therefore, the learning rate was set to 0.01. For CNN, ResNet, and LSTM, the learning rate was set to 0.0001. The number of epochs for each model was set to 200. Network training loss and accuracy are shown in Fig. 10 [Figure 10: see original paper]. All models converged quickly (training loss decreased exponentially and steadily, reaching approximately 0.1 within 50 epochs). The final nuclide identification accuracy for each model is shown in Table 4 . Overall, the accuracy of identification methods based on neural networks exceeded 90%. BP has a simple structure and can stabilize within 50 epochs; however, it easily reaches local optima, preventing further accuracy improvement. The "dropout layer" can effectively alleviate BP overfitting, with the worst accuracy rate for BP without dropout after training being 83%. This demonstrates BP's tendency to fall into local optima. After adding a dropout layer with a dropout rate of 0.2, BP's convergence accuracy reached 92.8%. CNN and LSTM networks had better global optimization capabilities than BP, with all three training methods converging stably at approximately 95%. Compared with CNN, RNN has the disadvantage of non-parallelizable computation, which is an important factor causing its longer training cycle. The step length of LSTM (based on RNN) is longer than that of RNN, and corresponding forward and backward propagation steps are also increased, resulting in greater training resource consumption. Therefore, LSTM requires a longer training cycle. ResNet, as an improved CNN model, allows deeper convolutions to be fully trained. Its ability to extract data features is better than that of CNN, and the overall recognition rate increased by approximately 1%.

The TBNN model proposed in this paper has a hyperparameter magnitude similar to that of ResNet, both on the order of millions. Within 20 training cycles, the recognition rate is approximately 83%. However, as the number of training cycles increased, the potential of TBNN was revealed. The optimal recognition rate reached 98.7%, which is at least 2.4% higher than that of the other models. This indicates that the transformer network structure paradigm has significant potential for nuclide identification.

Table 4. Recognition rates of each model (averaged over three training runs)

Model	Single Nuclide Accuracy	Total Nuclide Accuracy
BP	99.1%	92.8%
CNN	99.3%	95.3%
ResNet	99.4%	96.3%
LSTM	99.3%	96.6%
TBNN	99.3%	98.7%

Single nuclide accuracy refers to the recognition rate for individual nuclide types, while total nuclide accuracy refers to correct recognition of all nuclides present in a spectrum.

IV. Conclusion

This paper proposes, for the first time, a nuclide identification model based on the transformer architecture. Unlike traditional neural network architectures such as BP and CNN, this study explores the potential of the encoder-decoder paradigm based on the self-attention mechanism for artificial intelligence-based nuclide identification. In the NLP field, large language models have reached hundreds of millions of hyperparameters. Since transformer input is typically sequential, this study converts the 1024-channel one-dimensional spectrum into a 32×32 spectrum sequence as TBNN input. Experimental verification demonstrated that this spectrum-to-sequence conversion is an effective processing method that retains all spectrum data without generating excessive model nodes, enabling training under limited graphics card resources.

We established a database for eight industrial radioactive nuclides (^{241}Am, ^{192}Ir, ^{226}Ra, ^{133}Ba, ^{60}Co, ^{57}Co, ^{137}Cs, and ^{152}Eu) and constructed four representative neural network models (BP, CNN, ResNet, and LSTM) to validate the proposed new model. When the number of attention heads was four and the number of layers was four, the TBNN model achieved the best nuclide identification performance on the established dataset, with an identification rate of 98.7%. Based on comparison results, we infer that the TBNN's ability to extract data features is not inferior to traditional neural network paradigms. It is also noteworthy that in the 8-nuclide recognition task, TBNN with four attention heads and four network layers achieved optimal results, and increasing attention heads and network layers did not significantly reduce recognition rate. This suggests that in more complex tasks, such as nuclide identification on databases with richer nuclide types, deeper TBNN models will remain highly competitive, representing a direction for future research. Overall, our research demonstrates the effectiveness of introducing transformer models into the field of artificial intelligence for nuclide identification.

Acknowledgements

We thank the Sichuan Provincial Key Laboratory of Applied Nuclear Techniques in Geosciences and CDUT Team 203 for their English language review. We also

thank Professors Liang-Quan Ge, David Cohen, Sheng-Qing Xiong, and Si-Chun Zhou for their support.

Author Contributions

All authors contributed to the study conception and design. Algorithm proposal and verification, raw data generation, and analysis were performed by Fei Li, Chu-Yang Luo, Feng Cheng, Guo-Qiang Zeng, and Sheng Lv. The first draft of the manuscript was written by Fei Li, Bing-Hai Li, Jian-Feng Jiang, and Ying-Zi Wen, and all authors commented on previous versions of the manuscript. All authors have read and approved the final version of the manuscript.

Data Availability

Data will be made available on request.

References

[1] X. Li, C. Dong, Q. Zhang et al., Research and design of a rapid nuclide recognition system. J. Instrum. 17, T06008 (2022). doi:10.1088/1748-0221/17/06/T06008
[2] A.W. Ajlouni, M.M. Alnairi, K.S. Albarkaty et al., Nuclear security in public events. J. Radiat. Res. Appl. Sci. 16, 100572 (2023). doi:10.1016/j.jrras.2023.100572
[3] H. Xu, X. Ai, Y. Wang et al., Ground radioactivity distribution reconstruction and dose rate estimation based on spectrum deconvolution. Sensors-basel. 23, 5628 (2023). doi:10.3390/s23125628
[4] S. Qi, W. Zhao, Y. Chen et al., Comparison of machine learning approaches for radioisotope identification using nai (ti) gamma-ray spectrum. Appl. Radiat. Isotopes 186, 110212 (2022). doi:10.1016/j.apradiso.2022.110212
[5] Jozsef Suto and Stefan Oniga, Efficiency investigation from shallow to deep neural network techniques in human activity recognition. Cognitive Systems Research 54, 37–49 (2019). doi:10.1016/j.cogsys.2018.11.009
[6] F. Li, Z.X. Gu, L.Q. Ge, Review of recent gamma spectrum unfolding algorithms and their application. Results in Physics 13, 102211 (2019). doi:10.1016/j.rinp.2019.102211
[7] F. Li, Z.Y. Cheng, C.S. Tian, Progress in recent airborne gamma ray spectrometry measurement technology. Applied Spectroscopy Reviews 56(4), 255–288 (2021). doi:10.1080/05704928.2020.1768107
[8] X.Z. Li, Q.X. Zhang, H.Y. Tan et al., Fast nuclide identification based on a sequential bayesian method. Nucl. Sci. Tech. 32, 143 (2021). doi:10.1007/s41365-021-00982-z
[9] Z.J. Yang, W.B. Wang, P. Huang et al., Local weighted linear representation classifier based on regression recognition. Comput. Sci. 48, 351–359 (2021). doi:10.11896/jsjkx.210100173
[10] J.M. Wang, C-Land Zhang, Rapid detection of weak radionuclide in moving target. Information Technology and Network Security 37, 137–139 (2018). doi:10.19358/j.issn.2096-5133
[11] Y.G. Huo, H.Z. An, Y. L.Wu et al., Fast radioactive nuclide recognition

method study based on pattern recognition. Nuclear Electronics Detection Technology 34, 51–53 (2014). doi:10.3969/j.issn.0258-0934.2014.01.013
[12] H.L. Liu, The study of radioactive nuclide identification method based on fuzzy decision tree., https://d.wanfangdata.com.cn/thesis/D01476997 (2018)
[13] X.B. Xie, H. Zhang, The study of fast nuclide identification method with svm. Computer Knowledge and Technology 355–358 (2014). https://d.wanfangdata.com.cn/periodical/dnzsyjs-itrzyksb201402044
[14] J.M. Zhang, H.B. Ji, X.H. Feng et al., Nuclide spectrum feature extraction and nuclide identification based on sparse representation. High Power Laser and Particle Beams 30, 046003–1 (2018). doi:10.11884/hplpb201830.170435
[15] J.S. Ren, J.M. Zhang, K.P. Wang, Radioactive nuclide identification method based on svd and svm. Ordnance Industry Automation 36, 50–53 (2017). doi:10.7690/bgzdh.2017.05.014
[16] S. Zhou, Research on nuclide recognition method based on fuzzy decision tree with multi-category weight label. Master's thesis, Southwest University of Science and Technology, https://d.wanfangdata.com.cn/thesis/D02661842 (2022)
[17] G. Juan, The principle and simulation example of artificial neural network, (Beijing: China machine press, 2003), https://www.zhangqiaokeyan.com/book-cn/081503037434.html
[18] J. Wang, W. Gu, H. Yang et al., Analytical method for γ energy spectrum of radioactive waste drum based on deep neural network. Nucl. Tech. 45, 53–59 (2022). doi:10.11889/j.0253-3219.2022.hjs.45.040501
[19] N. He, H.Y. Lv, B. Wang et al., Nuclide identification method based on artificial neural network. Ordnance Industry Automation 3, 03 (2022). doi:10.7690/bgzdh.2022.03.018
[20] M.Y. Zhu, C.Q. Yuan, Y.F. Liu et al., Research on the fast recognition algorithm of nuclide based on bp neural network. Nuclear Electronics Detection Technology 38, 138–142 (2018). doi:10.3969/j.issn.0258-0934.2018.02.026
[21] Y.C. Liu, W. Wang, D.Q. Niu, Nuclide identification and analysis using artificial neural network. Ordnance Industry Automation 34, 86–91 (2015). doi:10.7690/bgzdh.2015.11.022
[22] J. He, X. Tang, P. Gong et al., Rapid radionuclide identification algorithm based on the discrete cosine transform and bp neural network. Ann. Nucl. Energy 112, 1–8 (2018). doi:10.1016/j.anucene.2017.09.032
[23] J. Kim, K. Park, G. Cho, Multi-radioisotope identification algorithm using an artificial neural network for plastic gamma spectra. Appl. Radiat. Isotopes 147, 83–90 (2019). doi:10.1016/j.apradiso.2019.01.005
[24] Y.C. Liu, H.G. Zhu, Y.Q. Song, Optimize bp neural network by an improved particle swarm optimization to implement nuclide identification. Ordnance Industry Automation 35, 88–92 (2016). doi:10.7690/bgzdh.2016.04.023
[25] K. He, X. Zhang, S. Ren et al., in Proceedings of the IEEE conference on computer vision and pattern recognition, Deep residual learning for image recognition. 2016, pp. 770–778. doi:10.48550/arXiv.1512.03385
[26] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning. Electron. Mark. 31, 685–695 (2021). doi:10.1007/s12525-021-00475-2

[27] H. Haohang, Z. Jiangmei, W. Kunpeng et al., Application of convolutional neural networks in identification of complex nuclides. Transducer and Microsystem Technologies 38, 154–156 (2019). doi:10.13873/J.1000-9787(2019)10-0154-03
[28] D. Liang, P. Gong, X. Tang et al., Rapid nuclide identification algorithm based on convolutional neural network. Ann. Nucl. Energy 133, 483–490 (2019). doi:10.1016/j.anucene.2019.05.051
[29] W. Yao, L. Zhiming, W. Yaping et al., Energy spectrum nuclide recognition method based on long short-term memory neural network. High Power Laser and Particle Beams 32, 106001–1 (2020). doi:10.11884/HPLPB202032.200118
[30] Ri Zhao and Na Liu, Low-resolution gamma-ray spectrum analysis using comprehensive training set and deep ResNet architecture. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 1050, 168135 (2023). doi:10.1016/j.nima.2023.168135
[31] J.Y. Guo, Q. Cai, J.P. An et al., A transformer based neural network for emotion recognition and visualizations of crucial eeg channels. Physica A 603, 127700 (2022). doi:10.1016/j.physa.2022.127700
[32] F. Alamri, A. Dutta, Implicit and explicit attention mechanisms for zero-shot learning. Neurocomputing 534, 55–66 (2023). doi:10.1016/j.neucom.2023.03.009
[33] M. Springenberg, A. Frommholz, M. Wenzel et al., From modern cnns to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Med. Image Anal. 87, 102809 (2023). doi:10.1016/j.media.2023.102809
[34] Y. Xia, Y. Xiong, K. Wang, A transformer model blended with cnn and denoising autoencoder for inter-patient ecg arrhythmia classification. Biomed. Signal Proces. 86, 105271 (2023). doi:10.1016/j.bspc.2023.105271
[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv e-prints arXiv:2010.11929 (2020). arXiv:2010.11929, doi:10.48550/arXiv.2010.11929
[36] C. Si, W. Yu, P. Zhou et al., Inception transformer. Adv. Neur. 35, 23495–23509 (2022). https://www.engineeringvillage.com/app/doc/?docid=cpx_{58dfe1131891d4db3b7M6ed310178
[37] R. Czarwinski, W. Weiss, Safety and security of radioactive sources–international provisions. Kerntechnik 70, 315–321 (2005). doi:10.3139/124.100262
[38] V. Smolyar, V. Tarasov, A. Mileva et al., Geant4 simulation of the moderating neutrons spectrum. Radiat. Phys. Chem. 212, 111151 (2023). doi:10.1016/j.radphyschem.2023.111151
[39] L. Zhe, L. Min, S. Rui et al., Monte carlo simulation and gaussian broaden techniques for full energy peak of characteristic x-ray in edxrf. Nucl. Tech. 35, 911–915 (2012). https://webofscience.clarivate.cn/wos/alldb/full-record/CSCD:4717606
[40] M.H. Guo, T.X. Xu, J.J. Liu et al., Attention mechanisms in computer vision: A survey. Computational visual media 8, 331–368 (2022). doi:10.1007/s41095-022-0271-y
[41] A. Vaswani, N. Shazeer, N. Parmar et al., Attention Is All You Need. arXiv e-prints arXiv:1706.03762 (2017). arXiv:1706.03762, doi:10.48550/arXiv.1706.03762

[42] C. Van Hiep, D.T. Hung, N.N. Anh et al., Nuclide identification algorithm for the large-size plastic detectors based on artificial neural network. IEEE Trans. Nucl. Sci. 69, 1203–1211 (2022). doi:10.1109/TNS.2022.3173371
[43] H.C. Lee, B.T. Koo, J.Y. Jeon et al., Radionuclide identification based on energy-weighted algorithm and machine learning applied to a multi-array plastic scintillator. Nucl. Eng. Technol. 55, 3907–3912 (2023). doi:10.1016/j.net.2023.07.005
[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90 (2017). doi:10.1145/3065386
[45] S.K. Sahu, R.N. Yadav, Key facial points recognition using resnet. Mater. Today:. Proc. 66, 3651–3656 (2022). doi:10.1016/j.matpr.2022.07.342
[46] D. Yilmaz, İ.E. Büyüktahtakın, in Operations Research Forum, Learning optimal solutions via an lstm-optimization framework. Vol. 4, Springer, 2023, p. 48. doi:10.48550/arXiv.2207.02937
[47] N. Pei, Y. Wu, R. Su et al., Interval prediction of the permeability of granite bodies in a high-level radioactive waste disposal site using lstm-rnns and probability distribution. Front. Earth Sci. 10, 835308 (2022). doi:10.3389/feart.2022.835308
[48] A. Neelakantan, T. Xu, R. Puri et al., Text and Code Embeddings by Contrastive Pre-Training. arXiv e-prints arXiv:2201.10005 (2022). arXiv:2201.10005, doi:10.48550/arXiv.2201.10005

Note: Figure translations are in progress. See original paper for figures.