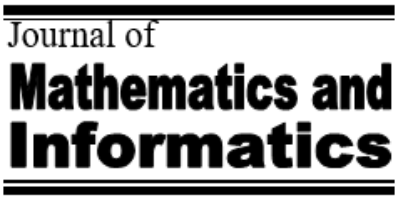
*Journal of Mathematics and Informatics*
*Vol. 30, 2026, 31-39*
*ISSN: 2349-0632 (P), 2349-0640 (online)*
*Published 30 April 2026*
www.researchmathsci.org
*DOI: http://dx.doi.org/10.22457/jmi.v30a04270*




# IFAN: An Interpretable Neural Network Combining Fuzzy Inference System with Attention Mechanism

***Yuan Liu*[1] *and Na Qin*[2*]**

[1,2]Department of Artificial Intelligence and Computer Science,
Northwest Normal University, Lanzhou, 730070, Lanzhou, China.
[1]Email:202421162029 @nwnu.edu.cn
[2*]Corresponding author. Email: super_qn@126.com

*Received 14 March 2026; accepted 28 April 2026*

***Abstract.*** To address the issues that attention weights are deterministic scalars lacking fine-grained characterization of feature importance, and that the weight computation process lacks interpretability, this paper presents an Interpretable Fuzzy Attention Network (IFAN), which combines fuzzy logic reasoning with the attention mechanism. The method fuzzifiers feature values into semantic concepts using learnable Gaussian membership functions, constructs a structured fuzzy rule base, performs inference using rule weights, and finally generates attention weights through defuzzification, achieving adaptive and interpretable multi-scale feature fusion. All fuzzy parameters are learnable end-to-end via backpropagation without manual rule design. Experiments on the CIFAR-10 dataset show that IFAN improves classification accuracy by 1.1 percentage points over a plain CNN and outperforms channel attention and spatial attention mechanisms. Furthermore, the internal interpretable decision basis of the model is provided from three aspects: membership functions, rule weights, and Local Interpretable Model-agnostic Explanations（LIME）attribution analysis.

***Keywords:*** Interpretable Fuzzy Attention Network; semantic interpretation; attention mechanism; fuzzy reasoning; gradient analysis

***AMS Mathematics Subject Classification (2010):*** 68T07

## 1. Introduction

Attention mechanisms enable deep neural networks to focus on task-relevant information by dynamically weighting input features. Since Bahdanau et al. [1] introduced attention into neural machine translation, various attention models, such as self-attention [2], channel attention [3], and spatial attention, have achieved great success in computer vision,

natural language processing, and other fields. However, existing attention mechanisms have two major limitations: (1) attention weights are typically given as deterministic scalars, lacking a fine-grained characterization of the "degree of importance"; (2) the computation process of attention weights is akin to a "black box", making it difficult to explain to users why a particular feature is assigned a specific weight.

Fuzzy inference systems [4], through membership functions and IF-THEN rules, simulate human approximate reasoning and naturally possess the ability to handle uncertainty and provide semantic explanations. In recent years, a few works have attempted to integrate fuzzy systems with deep learning [5], but most focus on fuzzy neural networks or fuzzy classifiers, and have not systematically established a theoretical framework for "fuzzy attention" from a mathematical perspective.

To address the above problems, we propose a fuzzy attention mechanism combined with fuzzy rules and provide a rigorous analysis of its mathematical properties. The main contributions are as follows:

1. Propose a novel fuzzy attention module that combines fuzzy logic with attention to achieve adaptive and interpretable multi-scale feature fusion.

2. Construct an end-to-end image classification network based on fuzzy attention, where all fuzzy parameters are learnable via backpropagation without manual rule design.

3. Provide interpretability analysis through mathematical methods, focusing on parameter semantics and attribution analysis.

## 2. Preliminaries

A fuzzy set is an extension of classical set theory [6]. In a classical set, an element either belongs to a set or does not; in a fuzzy set, an element can belong to the set to a certain degree, which is given by a membership function with values in [0,1]. For example, the brightness of a pixel can simultaneously belong to "bright" with a degree of 0.7 and to "medium brightness" with a degree of 0.2.

Fuzzy rules typically take the "IF-THEN" form, e.g., "IF the feature value is high THEN this channel is more important". The IF part is called the antecedent, and the THEN part is called the consequent. Through a set of rules, a fuzzy system can simulate human approximate reasoning [7].

Attention mechanism is a commonly used technique in deep learning. Its core idea is to assign different weights to different input features, so that the model focuses on the most beneficial information for the current task. For example, in image classification, the model may pay more attention to the edges or texture regions of the target object.

## 3. Interpretable fuzzy attention network.

### 3.1. Multi-branch feature extraction

Assume the input image $x$ is an RGB image of size 32×32×3,It passes through three parallel convolutional layers with kernel sizes 1×1,2×2 and 3×3, producing feature maps of size 32×32×6 each, the Global average pooling is applied to each branch to obtain global feature vectors, which are stacked into a tensor $G$ of size 3×6.

### 3.2. Fuzzification

For each branch and each channel, define $K = 3$ Gaussian membership functions， with centers $c_{ijk}$ and widths $w_{ijk}$ as learnable parameters. The centers are initialized using uniform random distribution in the range[0.1, 0.9]。The widths $w_{ijk}$ are initialized with a constant value of 0.15, and later constrained to remain within [0.05, 1.0] to maintain reasonable fuzzy set shapes. For a global feature value $g_{ij}$， the membership degree to the $k-th$ fuzzy set is computed as:

$$\mu_{ijk} = \exp\left(-\frac{1}{2}\left(\frac{g_{ij} - c_{ijk}}{w_{ijk}}\right)^2\right) \tag{1}$$

The output membership tensor $\mu$ ,has size3×6×3.

### 3.3. Fuzzy rule base

Stores learnable rule weights $R$ of size 3×6×3. Each weight corresponds to a rule: a specific combination of branch, channel, and fuzzy set. Initial values are uniformly random in[0.3, 0.7], During training, these rule weights are optimized via backpropagation. First, the rule activation strength is obtained by element-wise product of membership degrees and rule weights:

$$\alpha_{ijk} = \mu_{ijk} \cdot R_{ijk} \tag{2}$$

These activations then participate in subsequent inference and classification. The gradient of the classification loss with respect to the rule weights is automatically computed, and the Adam optimizer adjusts the weights to gradually reduce the loss function. To avoid overfitting and maintain semantic meaning, a two-stage training strategy is adopted: the first stage freezes the fuzzy layers and trains only the subsequent fully connected layers; the second stage unfreezes the fuzzy layers with a lower learning rate, allowing fine-tuning of rule weights based on learned features to better adapt to the classification task. After each parameter update, the weights are constrained by clipping back to the [0.1,1.0] interval, ensuring that the rule strengths remain within a reasonable range, thus preserving clear semantics and interpretability for each rule while improving classification performance.

### 3.4. Fuzzy inference

First, sum the activations over all channels and fuzzy sets within each branch to obtain the

overall activation strength for each branch:

$$\beta_i = \sum_{j=1}^{D}\sum_{k=1}^{K}\alpha_{ijk}, \quad i = 1,2,3 \tag{3}$$

Then, weight the branch activations using learnable branch aggregation weights $a_i$ :

$$\gamma_i = \beta_i \cdot a_i \tag{4}$$

The output $\gamma$ is a scalar (size 1×1×3 in the sense of branches)。To track branch importance, the inference layer records the average activation value per batch.

### 3.5. Defuzzification

Convert the inference result into normalized attention weights using the Softmax function:

$$w_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{3}\exp(\gamma_j)}, \quad i = 1,2,3 \tag{5}$$

The resulting fusion weights $w$ ,satisfy $\sum_{i=1}^{3} w_i = 1, \quad w_i > 0$ .

### 3.6. Weighted feature fusion

Expand the weights to the same dimensions as the branch feature maps and perform weighted summation:

$$\mathrm{F}_{\text{fused}} = \sum_{i=1}^{3} w_i \cdot \mathrm{F}_i \tag{6}$$

The fused feature map $\mathrm{F}_{\text{fused}}$ is then fed into the subsequent classification network.

## 4. Experimental validation and interpretability

### 4.1. Experimental setup

The dataset used is CIFAR-10. The baseline model is a plain CNN without attention, having the same architecture as IFAN but with the fuzzy attention module removed, directly convolving the original image. Both models use the Adam optimizer, an initial learning rate 0.001, batch size 128, and are trained for 40 epochs. Comparisons are also made with spatial attention and channel attention (SE).

### 4.2. Classification accuracy

**Table 1:** Classification performance comparison of different models on CIFAR-10

| Model | Accuracy(%) | Precision(%) | Recall(%) | F1-score(%) |
|---|---|---|---|---|
| Baseline CNN | 75.92 | 75.96 | 75.92 | 75.93 |
| SE | 76.75 | 76.78 | 76.75 | 76.76 |
| Spatial | 76.54 | 76.53 | 76.54 | 76.56 |
| IFAN | 77.02 | 77.00 | 77.02 | 77.00 |

As shown in Table 1, IFAN improves accuracy by 1.1 percentage points over the

baseline CNN, by 0.27 percentage points over SE, and by 0.48 percentage points over Spatial attention. Paired t-test p-value < 0.01, indicating statistical significance.

### 4.3. Parameter update

All learnable parameters of the fuzzy attention network are optimized end-to-end by minimizing the classification loss function. This section presents the loss function, analytical gradient expressions for each parameter, optimizer update rules, and training strategy.

#### 4.3.1. Loss function

The multi-class cross-entropy loss is used. Let the predicted probability vector for the $n-th$ sample be, $\mathrm{y}^{(n)} = \mathrm{Softmax}(\mathrm{o}^{(n)})$ where is the output of the last fully connected layer. The true label is the one-hot vector $y^{(n)}$. Then the loss function is:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{10} y_c^{(n)} \log \hat{y}_c^{(n)} \tag{7}$$

where, $N$ is the batch size.

#### 4.3.2. Analytical gradients

Based on the forward propagation of the fuzzy attention defined in Section 3, the chain rule yields gradients of the loss with respect to each layer's parameters. Let the gradient of the loss $L$ with respect to the fused feature map $\mathrm{F}_{\mathrm{fused}}$ be $\frac{\partial L}{\partial \mathrm{F}_{\mathrm{fused}}}$. Then, with respect to the attention weights $w_i$ after defuzzification:

$$\frac{\partial L}{\partial w_i} = \left\langle \frac{\partial L}{\partial \mathrm{F}_{\mathrm{fused}}}, \mathrm{F}_i \right\rangle \tag{8}$$

where $\langle *,* \rangle$ denotes the inner product.

With respect to the inference output $\gamma_i$ before defuzzification:

$$\frac{\partial L}{\partial \gamma_i} = \frac{\partial L}{\partial w_i} \cdot w_i - w_i \sum_{j=1}^{3} \frac{\partial L}{\partial w_j} w_j \tag{9}$$

With respect to the branch aggregation weights $a_i$ :

$$\frac{\partial L}{\partial a_i} = \frac{\partial L}{\partial \gamma_i} \cdot \beta_i \tag{10}$$

where, $\beta_i = \sum_{j=1}^{D}\sum_{k=1}^{K} \alpha_{ijk}$ is the raw branch activation strength.

With respect to the rule weights $R_{ijk}$ :

$$\frac{\partial L}{\partial R_{ijk}} = \frac{\partial L}{\partial \gamma_i} \cdot a_i \cdot \mu_{ijk} \tag{11}$$

With respect to the membership function centers $c_{ijk}$ and widths $w_{ijk}$ :

$$\frac{\partial L}{\partial c_{ijk}} = \frac{\partial L}{\partial \gamma_i} \cdot a_i \cdot R_{ijk} \cdot \mu_{ijk} \cdot \frac{g_{ij} - c_{ijk}}{\sigma_{ijk}^2} \tag{12}$$

$$\frac{\partial L}{\partial w_{ijk}} = \frac{\partial L}{\partial \gamma_i} \cdot a_i \cdot R_{ijk} \cdot \mu_{ijk} \cdot \frac{(g_{ij} - c_{ijk})^2}{w_{ijk}^3} \tag{13}$$

### 4.3.3. Adam optimizer

The Adam optimizer [6] is used to update parameters. Let the parameter at step $t$ be $\theta_t$, and the gradient be $g_t = \nabla_\theta L_t$. Then the update rules are:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{14}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{15}$$

$$m_t = \frac{m_t}{1-\beta_1^t}, \quad v_t = \frac{v_t}{1-\beta_2^t} \tag{16}$$

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t} + \dot{o}} \tag{17}$$

where $\eta$ is the learning rate, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\dot{o} = 10^{-7}$。

## 4.4. Interpretability analysis

Unlike black-box models that rely on external approximate explainers, all intermediate variables in the fuzzy attention mechanism have clear mathematical semantics, allowing direct extraction of interpretable decision bases from within the model. This section focuses on parameter semantics for interpretability analysis.

### 4.4.1. Semantic interpretability of membership functions

The center $c_{b,d,k}$ of each membership function

$$\mu_{b,d,k}(x) = \exp\left(-\frac{1}{2}\left(\frac{x - c_{b,d,k}}{w_{b,d,k}}\right)^2\right)$$

corresponds directly to a typical characteristic value. After training, these centers automatically cluster near the modes of the data distribution. For example, for branch 2, channel 3, if the three centers are $c \approx 0.25, 0.55, 0.78$ ,semantic labels can be assigned: low feature value (0.25), medium feature value (0.55), high feature value (0.78). The width $w_{b,d,k}$ reflects the coverage of the fuzzy set: smaller $w$ indicates a more precise concept; larger $w$ indicates a vaguer concept.

For any input feature value $g_d^{(b)}$ ,the membership vector $(\mu_{b,d,1}, \mu_{b,d,2}, \mu_{b,d,3})$ gives a soft assignment of the feature value to the three semantic concepts. For instance, if $\mu = (0.05, 0.92, 0.03)$ , the feature value is interpreted as "medium to high".

### 4.4.2. Semantic interpretation of rule weights

The rule weight $w_{b,d,k} \in [0,1]$ indicates "if the feature value of branch $b$ , channel $d$ belongs to fuzzy concept $k$ , then its contribution to the branch activation strength is $w_{b,d,k}$ ". After

training, larger $w_{b,d,k}$ indicates that the rule is trusted as important by the model. For example, if $w_{2,3,2} = 0.87$ while $w_{2,3,1} = 0.12$, the model considers "branch 2, channel 3 feature value medium" as strong evidence, while "low feature value" barely activates that branch. Define a rule importance metric:

$$\rho_{b,d,k} = w_{b,d,k} \cdot \mathrm{E}\left[\mu_{b,d,k}\left(g_d^{(b)}\right)\right] \tag{18}$$

where the expectation is estimated over the training set. This metric combines the intrinsic weight and the actual activation frequency of the rule, allowing ranking of the most important rules. For example, the top three rules might be:

$\rho_{2,3,2} = 0.71$ (branch 2, channel 3, medium value)

$\rho_{1,4,1} = 0.58$ (branch 1, channel 4, low value)

$\rho_{3,1,3} = 0.52$ (branch 3, channel 1, high value)

These rules provide human-readable explanations in natural language.

### 4.4.3. LIME attribution analysis

LIME generates perturbed samples and trains a local interpretable model to approximate the decision boundary of the original model [9]. For images, LIME segments the input into superpixels, perturbs combinations of superpixels, observes prediction changes, and evaluates the importance of each superpixel.

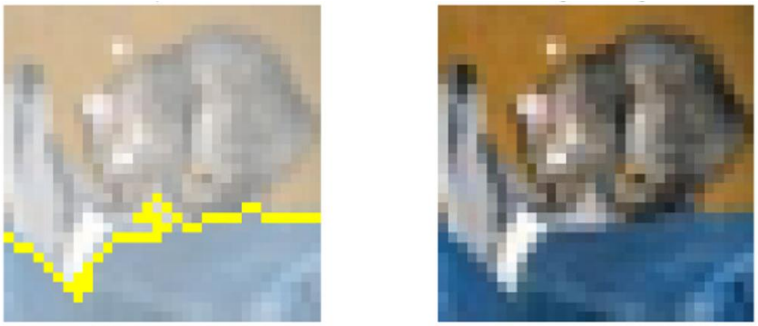


**Figure 1:** LIME Attribution Analysis

Figure 1 presents the LIME-based local interpretability visualization: the left image is the explanation map, and the right image is the original input. The yellow zero-value line serves as the core decision baseline. Regions above this line provide positive contributions, supporting the target category, whereas regions below it corresponds to negative contributions, which reduce prediction confidence. By mapping complex decisions to sparse feature contributions via a local linear surrogate model, LIME reveals both critical evidence and interfering factors. Consequently, it offers a verifiable explanation in terms of granularity, direction, and importance, thereby enhancing model transparency and trustworthiness.

### 4.4.4. Quantitative interpretability metrics

In addition to the qualitative semantic analysis and local attribution visualization above, this subsection introduces two quantitative metrics to objectively compare the interpretability of different attention mechanisms: entropy of attention weights and SHAP stability. The former measures the concentration of attention weight distribution, while the latter evaluates the robustness of the explanation to input perturbations. Table 2 presents the results of the above interpretability metrics for the three attention mechanisms on the CIFAR-10 test set.

**Table 2:** Comparison of attention weight entropy and SHAP stability

| Model | Entropy of Attention Weights | SHAP Stability |
|---|---|---|
| IFAN | 0.7342 | 0.0294 |
| SE | 2.7694 | 0.0337 |
| Spatial | 6.9295 | 0.0338 |

The entropy of Fuzzy Attentions much lower than that of Channel Attention and Spatial Attention. This indicates that the proposed fuzzy attention mechanism produces highly concentrated attention weights; the model can focus on a few critical branches or features with strong selectivity, which is consistent with the interpretable rules observed in Sections 4.4.1 and 4.4.2. In contrast, the entropy of Spatial Attention is nearly 7, implying that its weights are distributed over about $2^7 = 128$ equally probable units, essentially degenerating into a uniform distribution and losing selective attention.

Fuzzy Attention achieves a SHAP stability of 0.0294, outperforming the two baseline models. The lower stability value suggests that the explanation of the fuzzy attention mechanism is more robust to small input noise. This benefits from the soft interval partitioning of feature values by the fuzzy membership functions and the regularization constraints on rule weights, which smooth the decision boundary.

## 5. Conclusion

This paper proposes an Interpretable Fuzzy Attention Network (IFAN) that integrates fuzzy logic with attention mechanisms, achieving adaptive and interpretable multi-scale feature fusion through learnable Gaussian membership functions and fuzzy rule bases. Experiments on CIFAR-10 demonstrate that IFAN achieves superior classification accuracy compared to baseline CNN and standard attention models, with all fuzzy parameters optimizable end-to-end. The semantic transparency of membership functions and rule weights provides interpretable decision basis and attribution analysis. Future work will explore the generalization capability of fuzzy attention on Transformer architectures and larger-scale datasets [8].

***Acknowledgements.*** The authors sincerely thank the anonymous reviewers for their

valuable suggestions. This work is supported by the National Natural Science Foundation of China (Grant Nos. 12471444 and 12161082) and the Young Teachers' Scientific Research Ability Promotion Program of Northwest Normal University (Grant No. NWNU-LKQN2022-03).

***Conflict of Interest:*** The authors declare that they have no conflicts of interest.

***Authors' Contributions:*** Both authors contributed equally to this work.

## REFERENCES

1. A. Galassi, M. Lippi and P. Torroni, Attention in natural language processing, *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
2. S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan and M. Shah, Transformers in vision: a survey, *ACM Computing Surveys*, 54(10) Article 200, (2022) 1-41.
3. S. Mishra and R. K. Gupta, A comprehensive review of channel attention mechanisms in computer vision, *Neurocomputing*, 578 (2024) 127428.
4. M. H. Fazel Zarandi and S. Zarinbal, Thirty years after: new heuristic methods for designing fuzzy inference systems - a 2025 survey, *IEEE Transactions on Fuzzy Systems*, 33(2) (2025) 456-472.
5. E. Lughofer and M. Pratama, Evolving fuzzy neural networks for online streaming data learning: a 2024 perspective, *IEEE Transactions on Fuzzy Systems*, 32(6) (2024) 3456-3470.
6. Yuanyuan Zhang and Zengtai Gong, An approximate technique for solving fully fuzzy complex linear systems, *Journal of Mathematics and Informatics*, 28 (2025) 1-19.
7. V.R. Kulli, Reduced augmented Sombor index of certain networks, *Journal of Mathematics and Informatics*, 30 (2026) 1-5.
8. R. Wang and X. Jiang, Convolution pyramid attention: an efficient channel attention mechanism, *Proceedings of the Fourth International Conference on Computer Vision and Data Mining (ICCVDM 2023)*, p. 25, 2024.
9. A.M. Salih, Z. Raisi, I. Boscolo Galazzo, P. Radeva, S. Petersen, K. Lekadir, and G. Menegaz, A perspective on explainable artificial intelligence methods: SHAP and LIME, *Advanced Intelligent Systems*, 7 (2024) 2400304.
10. Taiwo O. Sangodapo, Homomorphism of picture fuzzy subgroup of a group, *Journal of Mathematics and Informatics*, 30 (2026) 7-16.