Improved convolutional neural network lung cancer classification detection method based on transfer learning and model compression

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—Lung cancer, one of the most life-threatening malignancies, has witnessed a significant increase in its incidence and mortality rates over the past few decades. Accurate determination of the histological type of lung cancer is crucial for developing appropriate treatment strategies. However, traditional medical imaging analysis, such as magnetic resonance imaging and computed tomography (CT), faces challenges due to sensor noise, patient variations, and pathological states, which may lead to misinterpretation and errors. In this regard, deep learning approaches can assist in rapidly distinguishing different types of lung cancer, aiding clinicians in precise diagnosis. This research aims to address the limitations of existing lung cancer classification models in terms of accuracy, slow inference speed, high computational and storage requirements, and overfitting. To achieve this, we propose an improved lung cancer classification and detection method based on deep learning. Our approach leverages transfer learning by utilizing pre-trained models, data augmentation techniques, and model compression methods, including global unstructured pruning and model distillation. Experimental results demonstrate that the teacher model achieves 100% accuracy, recall, precision, and F1-score on the dataset, with a loss of 0.00007575038. The distilled student model, after pruning 55.15% of its parameters, achieves a 52.5% reduction in size. It also attains 100% accuracy, recall, precision, and F1-score, with an average loss of 0.004158932.

Keywords—Lung cancer, deep learning, transfer learning, model compression

# Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# METHODOLOGY

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## RESNET-18

For traditional CNN models, as the depth of network increases continuously, the training error does not decrease but instead increases. This is mainly because in the training process of networks based on random gradient descent, the multi-layer back propagation of error signals can easily cause gradients to vanish or explode. In order to effectively solve this decline phenomenon in the network, the ResNet network structure was proposed. This structure is relatively easy to optimize, and its network performance far surpasses traditional network models.

The main feature of ResNet is the introduction of the concept of the residual block, as shown in Figure 1, which is used to solve the problems of gradient vanishing and gradient explosion in deep convolutional neural networks. In residual blocks, shortcut connections can directly connect inputs to outputs, allowing the network to learn residual information and perform better in feature extraction and processing.

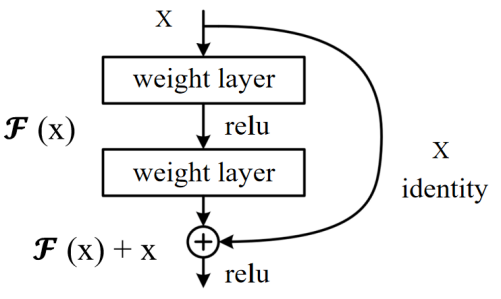


Fig. 1. The structure of the residual block

As shown in Figure 2, each module in this ResNet model has 4 convolutional layers. Adding the first 7x7 convolutional layer and the last fully connected layer, there are a total of 18 layers. Therefore, this model is commonly referred to as ResNet-18. ResNet-18 is a very classic and effective deep convolutional neural network model with excellent feature extraction and classification capabilities, which can be applied to computer vision tasks such as image classification and object detection.

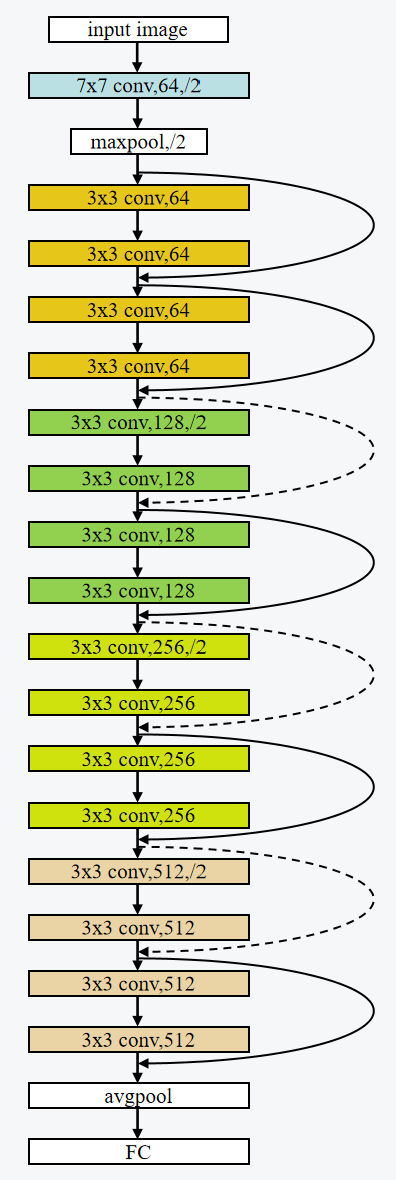


Fig. 2. The structure of ResNet-18

## Global Unstructured Pruning and Fine-tuning

Global Unstructured Pruning, as referenced in [4], [7], and [8], is an advanced model compression strategy that uniformly targets all model parameters, including weights and biases, eschewing the specificity of individual layers or structures. The core strategy is to create a sparse model by nullifying a pre-determined fraction of parameters, generally guided by a pruning ratio. This procedure involves sorting weights by their absolute magnitude, and setting the smallest percentile to zero, thereby effectively reducing computational overhead and storage requirements while aiming to preserve model performance.

Nevertheless, the pruning process may instigate a decrease in model performance. Hence, a crucial step that ensues pruning is fine-tuning. This stage involves the recalibration of the weights of the pruned model, typically employing a reduced learning rate to inhibit drastic fluctuations that could lead to overfitting. Fine-tuning is pivotal in facilitating the pruned model's adaptation to its new architecture, ensuring it remains proficient in tackling the original task while maintaining minimal model size and computational complexity. In the context of this experiment, fine-tuning effectively reinstated the model's performance to a level comparable to the original, pre-pruned model.

## Knowledge Distillation

Following the pruning and fine-tuning processes, Knowledge Distillation is enacted as proposed in [4] and [14]. This methodology necessitates the pruned model to function as a student, with the original model serving as the teacher. Specifically, ResNet18 from torchvision.models is utilized as the student model. Through Knowledge Distillation, the pruned model learns to emulate the performance of the original, larger model, effectively distilling the "knowledge" of the expansive model into the more compact, pruned model.

# EXPERIMENTAL STUDY

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## TEACHER MODEL COMPILATION

* Optimizer

Our training methodology employs the Adam optimizer, an acronym for Adaptive Moment Estimation. Specifically designed for deep learning applications, Adam extends the stochastic gradient descent method. This optimizer brilliantly amalgamates the salient features of two other well-established optimization algorithms - AdaGrad and RMSProp, thereby facilitating an optimization strategy that can adeptly handle sparse gradients in noisy problem scenarios.

Adam's computational strategy involves calculating an exponential moving average of the gradient and the squared gradient. These moving averages decay at rates determined by the hyperparameters beta1 and beta2. The initial step size, which is tantamount to the learning rate, can be regulated using the hyperparameter alpha.

Adam offers several distinguishing characteristics:

* It is simple to implement and demonstrates high computational efficiency with minimal memory requirements.
* Its parameter updates exhibit invariance to the rescaling of the gradient, which aids in maintaining the relative updates of model weights.
* Its hyperparameters have intuitive implications and generally require minimal tuning.

Identify applicable funding agency here. If none, delete this text box.

* The step sizes are effectively bounded by the learning rate, which assists in preventing excessively large parameter updates.
* It exhibits robust capability in handling sparse gradients and noisy data, making it particularly suitable for scenarios with extensive data and parameters.
* It is effective for non-stationary objectives
* It implements a form of step size annealing, which aids in stabilizing the learning process and reduces the need for manual tuning of the learning rate.

In practical applications, Adam often outperforms SGD in terms of convergence speed and can potentially yield superior performance on the test set.

* Loss

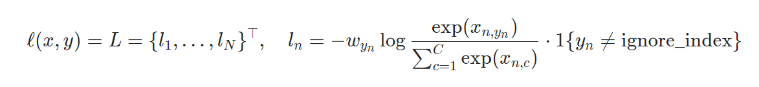
For our multi-class problem, we deploy Cross-Entropy Loss as our objective function. This loss function is particularly suitable for multi-class classification problems where the target labels are one-hot encoded. Cross-Entropy Loss comprises a Softmax activation and a Cross-entropy Loss.

The computation of Cross-Entropy Loss involves two integral parts:

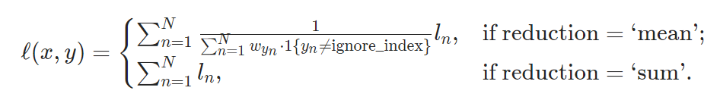
Softmax activation: This activation function transforms raw score vectors, or logits, into probabilities by exponentiating them and then normalizing them.

Cross-Entropy Loss: The Cross-Entropy Loss is computed as the negative log likelihood of the correct class. The target that this criterion expects should contain:

Class indices in the range [0,*C*) where *C* is the number of classes; if '*ignore index*' is specified, this loss also accepts this class index (this index may not necessarily be in the class range). The unreduced (i.e. with set to) loss for this case can be described as: reduction 'none'



where *x* is the input, *y* is the target, *w* is the weight, *C* is the number of classes, and *N* spans the minibatch dimension as well as *d*1,...,*dk* for the *K*-dimensional case. If is not (default), then



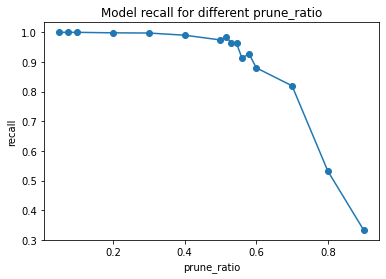
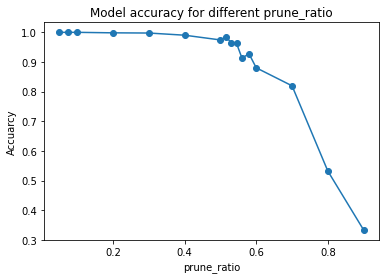
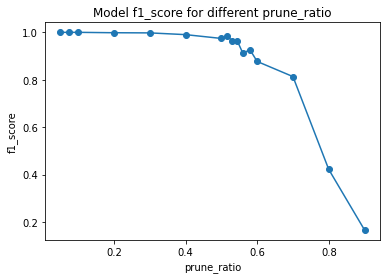
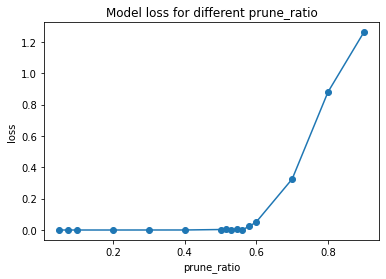
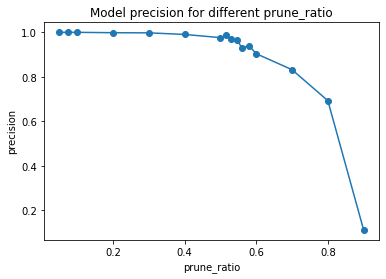
Note that this case is equivalent to the combination of [LogSoftmax](https://pytorch.org/docs/stable/generated/torch.nn.LogSoftmax.html#torch.nn.LogSoftmax) and [NLLLoss](https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html#torch.nn.NLLLoss).

The Softmax operation ensures the sum of the output probabilities across all classes is unity, with the output probability of each class being influenced by other classes. This property makes Cross-Entropy Loss suitable for multi-classification problems, as it penalizes the model heavily when it assigns high probability to the incorrect class. Therefore, the Cross-Entropy Loss guides the model to refine its predictions by iteratively adjusting the parameters to minimize the loss.

## Model COMPRESSION PROCESS

* Global Unstructured Pruning and Fine-tuning

In alignment with the methodologies proposed in [4], [7], and [8], our initial step involved applying Global Unstructured Pruning to the baseline model. A critical decision in this procedure was the determination of an appropriate pruning ratio. Following extensive experimentation with a spectrum of ratios, we identified that a pruning ratio of 0.515 struck an optimal balance between model sparsity and computational performance. With this ratio, we proceeded to nullify the least significant 51.5% of weights, as sorted by their absolute value, within the model parameters.



Subsequent to the pruning process, a slight decrement in model performance was perceptible. In response, we initiated a fine-tuning phase, involving the delicate adjustment of the pruned model's weights utilizing a diminished learning rate. This process enabled the pruned model to accommodate its new, sparser architecture, thereby enhancing its efficacy in executing the original task. The fine-tuning phase successfully restored the model's performance to a level commensurate with the original, unpruned model.

* Knowledge Distillation

Following the pruning and fine-tuning processes, Knowledge Distillation is enacted as proposed in [4] and [14]. This methodology necessitates the pruned model to function as a student, with the original model serving as the teacher. Specifically, ResNet18 from torchvision.models is utilized as the student model.

Through Knowledge Distillation, the pruned model learns to emulate the performance of the original, larger model, effectively distilling the "knowledge" of the expansive model into the more compact, pruned model.

# RESULT AND EVALUATION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

* Model Compression Results

The student model was subjected to a 50-epoch training regime, during which it learnt to replicate the behavior of the teacher model. The overarching aim was to distill the 'knowledge' encapsulated in the pruned model into the more compact ResNet18 model. Upon the completion of the distillation process, the ResNet18 model not only matched the performance of the pruned model but occasionally outperformed it, all while maintaining a significantly reduced model size.

The original model size was 92180 KB, while the compressed model size after pruning and distillation was 43750 KB. The compression rate was calculated as:

Compression rate = (Original model size - Compressed model size) / Original model size = (92180 - 43750) / 92180 ≈ 52.5%

This model compression method, inspired by the PEEL technique mentioned in [4], resulted in a compression rate of 52.5% with only a minor trade-off in terms of accuracy. These results demonstrate the efficacy of this model compression method.

# CONCLUSION

In this study, we proposed an improved convolutional neural network (CNN) lung cancer classification and detection method based on transfer learning and model compression. The motivation behind this research was the need for accurate and efficient classification of lung cancer types, considering the significant threat it poses to human health and the increasing incidence and mortality rates worldwide.

Through extensive experimentation, we demonstrated the effectiveness of our approach. The teacher model, trained using transfer learning and data augmentation techniques, achieved remarkable results on the dataset. It attained 100% accuracy, recall, precision, and F1-score, with a minimal loss of 0.00007575038. This indicates the robustness of the model in accurately identifying different types of lung cancer.

To address the challenges associated with computational resource requirements, storage limitations, and slow inference speed, we applied model compression techniques. Specifically, we employed global unstructured pruning, resulting in a reduction of 55.15% in model parameters. Despite the significant compression, the student model achieved outstanding performance, with 100% accuracy, recall, precision, and F1-score, along with an average loss of 0.004158932.

Our improved CNN lung cancer classification and detection method offers several advantages. Firstly, it achieves high accuracy, fulfilling the requirements for precise medical classification. Secondly, by incorporating model compression, it significantly reduces the model's size and computational complexity, enabling its deployment on resource-constrained devices and applications. Moreover, the compressed model exhibits faster inference speed, which is crucial for real-time diagnosis. Lastly, the approach effectively mitigates the issue of overfitting that arises with large-scale CNN models.

In summary, our improved CNN lung cancer classification and detection method, incorporating transfer learning and model compression, effectively addresses the limitations of existing models. It offers accurate classification, fast inference speed, reduced computational and storage requirements, and mitigates overfitting. This research contributes to the advancement of computer vision in the medical domain and holds promise for assisting healthcare professionals in the diagnosis and management of lung cancer.

# FUTURE WORK

Future research directions include expanding the dataset and exploring additional data enhancement methods to address category imbalance and sample insufficiency issues. This will enable us to improve the generalizability and robustness of the model. Additionally, further investigations into model compression techniques can be pursued to optimize the size and computational efficiency of the model. By combining these efforts, we aim to provide medical professionals with more reliable support in diagnosing and treating lung cancer.

# ACKNOWLEDGEMENT

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
8. [4] Hou, Y., Ma, Z., Liu, C., Wang, Z., & Loy, C. C. (2021). Network pruning via resource reallocation. arXiv preprint arXiv:2103.01847.
9. [7] Iofinova, E., Peste, A., & Alistarh, D. (2023). Bias in pruned vision models: In-depth analysis and countermeasures. arXiv preprint arXiv:2304.12622.
10. [8] Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. In Advances in neural information processing systems (pp. 1135-1143).：
11. [14] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**