1. Parameter-Efficient Fine-Tuning (PEFT):
Definition: A technique in machine learning that updates only a small subset of parameters in large models to reduce computational costs and improve efficiency.
Key points:
• Allows for effective adaptation of large pre-trained models to new tasks without retraining all parameters.
• Reduces memory and computational requirements, making it feasible to deploy large models in resource-constrained environments.
2. Sparse Fine-Tuning:
Definition: A method that selectively updates only a small number of parameters in a model, maintaining most parameters fixed to enhance efficiency.
Key points:
• Often employs techniques like pruning to identify which parameters to update, minimizing computational overhead.
• Helps maintain model performance while significantly reducing the number of trainable parameters.
3. Lottery Ticket Hypothesis:
Definition: A theory suggesting that within a randomly initialized neural network, there exist smaller subnetworks (winning tickets) that can be trained to achieve comparable performance.
Key points:
• Supports the idea that large models can be pruned to smaller, efficient models without losing accuracy.
• Provides a framework for identifying effective subnetworks that can be trained more quickly than full networks.
4. Low-Rank Adaptation (LoRA):
Definition: A technique that introduces low-rank matrices into the weight updates of a neural network, allowing for efficient fine-tuning with fewer parameters.
Key points:
• Reduces the number of parameters that need to be trained, leading to faster training times and lower memory usage.
• Enables the model to adapt to new tasks while maintaining the benefits of pre-trained weights.
5. Adapter Functions:
Definition: Small, trainable modules inserted into pre-trained models to adapt them for specific tasks without extensive modifications to the original model.
Key points:
• Allows for task-specific adaptations while preserving the integrity of the original model architecture.
• Facilitates quick adaptation to new tasks by adding minimal additional parameters.
6. Prompting:
Definition: A method where specific input phrases (prompts) guide the model's predictions, often used in large language models to generate contextually relevant responses.
Key points:
• Enables models to perform tasks without retraining by providing context through carefully crafted prompts.
• Can lead to variability in performance based on prompt wording, requiring careful design for optimal results.
7. Pruning:
Definition: A technique that removes unnecessary weights from a trained neural network to create a smaller, more efficient model while maintaining performance.
Key points:
• Can lead to significant reductions in model size and inference time, making models more deployable.
• Often involves criteria such as weight magnitude to determine which connections to remove.
8. In-Context Learning:
Definition: A method where the model learns to perform tasks based on examples provided in the input, without explicit fine-tuning or retraining.
Key points:
• Allows for rapid adaptation to new tasks by leveraging examples in the input context.
• Reduces the need for extensive labeled datasets for every new task, enhancing flexibility.
9. Catastrophic Forgetting:
Definition: A phenomenon where a neural network forgets previously learned information upon learning new information, particularly when fine-tuning on new tasks.
Key points:
• Can significantly hinder the performance of models that are continuously updated with new data.
• Strategies like modular representations can help mitigate this issue by preserving older knowledge.
10. Transfer Learning:
Definition: A machine learning approach where a model trained on one task is adapted for a different but related task, leveraging learned features.
Key points:
• Reduces the amount of data and time required to train models for new tasks.
• Commonly used in NLP to adapt models like BERT and GPT to various downstream tasks.
11. Fine-Tuning:
Definition: The process of adjusting a pre-trained model's parameters on a specific task to improve its performance on that task.
Key points:
• Involves training the model on a smaller dataset related to the target task, refining its predictions.
• Can lead to significant improvements in accuracy, especially for complex tasks.
12. Weight Magnitude Pruning:
Definition: A pruning technique that removes the smallest weights in a neural network, based on their magnitude, to create a sparser model.
Key points:
• Effective in reducing the model size without significantly impacting accuracy.
• Simplifies the model by focusing on the most impactful connections.
13. Bias-only Fine-Tuning (BitFit):
Definition: A fine-tuning approach that only updates bias parameters in a model while keeping all other weights fixed.
Key points:
• Requires minimal additional parameters, making it efficient for quick adaptations.
• Often achieves competitive performance with significantly fewer trainable parameters compared to full fine-tuning.
14. Structured Composition:
Definition: A method that imposes a structure on the selection of parameters in a model, allowing for more controlled updates during fine-tuning.
Key points:
• Can lead to better performance by ensuring that updates are coherent with the model's architecture.
• Useful in scenarios where certain layers or components are more relevant to the task at hand.
15. Multi-Layer Prompt Tuning:
Definition: A technique that allows for learning prompt parameters at multiple layers of a model, enhancing its ability to adapt to various tasks.
Key points:
• Increases the model's flexibility by allowing prompts to influence not just the input but also intermediate activations.
• Can improve performance on complex tasks where simple prompting may fall short.
1. Pretrained Language Models:
Definition: Models that have been trained on large text datasets to understand language patterns, enabling them to generate or classify text effectively.
Key points:
• Examples include BERT, GPT, and T5, which excel in various NLP tasks.
• They leverage transfer learning, allowing them to adapt to specific tasks with minimal additional training.
2. Decoding Methods:
Definition: Techniques used to generate text from language models, determining how the next word is selected based on probability distributions.
Key points:
• Common methods include greedy decoding, beam search, and sampling techniques.
• The choice of decoding method affects the quality and coherence of generated text.
3. Encoder-Decoder Architecture:
Definition: A neural network framework where one part (encoder) processes input data and another part (decoder) generates output, commonly used in translation tasks.
Key points:
• Used in models like BART and T5 for tasks such as summarization and translation.
• The encoder compresses input information, while the decoder reconstructs outputs based on this compressed data.
4. Masked Language Modeling:
Definition: A training technique where some words in a sentence are masked, and the model learns to predict these missing words from the context.
Key points:
• Employed in models like BERT to create bidirectional context, enhancing understanding of language.
• Helps in fine-tuning models for specific tasks like sentiment analysis or question answering.
5. Greedy Decoding:
Definition: A decoding strategy that selects the most probable word at each step, without considering future context, leading to quick but potentially suboptimal results.
Key points:
• Simple and efficient, but often produces generic or repetitive outputs.
• Suitable for tasks where speed is prioritized over creativity or diversity.
6. Beam Search:
Definition: A decoding method that maintains multiple candidate sequences at each step, selecting the most likely sequences based on cumulative probabilities.
Key points:
• Balances exploration and exploitation, improving text quality compared to greedy decoding.
• Can lead to higher computational costs due to maintaining multiple sequences.
7. Nucleus Sampling:
Definition: A sampling technique that focuses on a subset of the vocabulary, selecting from the top p portion of the probability distribution, enhancing diversity.
Key points:
• Reduces the influence of low-probability, irrelevant tokens, improving coherence.
• Effective for generating creative and contextually relevant text outputs.
8. Top-k Sampling:
Definition: A sampling method that selects the next word from the top k most probable candidates, ensuring that only high-probability options are considered.
Key points:
• Helps in generating diverse outputs while avoiding low-quality text.
• The choice of k impacts the balance between randomness and coherence in generated text.
9. Zero-shot Learning:
Definition: A machine learning paradigm where a model performs tasks without any task-specific training data, relying on its pretrained knowledge.
Key points:
• Demonstrated by models like GPT-2, which can perform various tasks based on context alone.
• Enables flexible application of models across different tasks without extensive retraining.
10. Transfer Learning:
Definition: A technique where knowledge gained from training on one task is applied to a different but related task, improving performance and reducing training time.
Key points:
• Essential for fine-tuning pretrained models on specific tasks with limited data.
• Facilitates rapid deployment of models across various applications in NLP.
11. Perplexity:
Definition: A measurement of how well a probability model predicts a sample, calculated as the exponentiation of the average negative log-likelihood of the predicted probabilities.
Key points:
• Lower perplexity indicates better predictive performance, reflecting how well the model understands the language.
• Used to compare the effectiveness of different language models and decoding methods.
12. Training Dataset Quality:
Definition: The importance of using high-quality, relevant datasets for training language models, impacting their performance and generalization capabilities.
Key points:
• High-quality datasets lead to better model performance and more coherent outputs.
• Models trained on noisy data may struggle with generating relevant and contextually appropriate text.
13. Fine-tuning:
Definition: The process of taking a pretrained model and further training it on a smaller, task-specific dataset to improve performance for that task.
Key points:
• Allows models to adapt to specific domains or tasks, enhancing their utility.
• Requires fewer labeled examples compared to training from scratch, saving time and resources.
14. Bidirectional Attention:
Definition: A mechanism in models like BERT that allows the model to consider context from both directions (left and right) when processing text, improving understanding.
Key points:
• Enhances the model's ability to capture nuanced meanings and relationships in text.
• Essential for tasks requiring deep contextual understanding, such as sentiment analysis.
15. Scaling Laws:
Definition: Observations in machine learning indicating that model performance improves predictably with increases in model size, dataset size, and compute resources.
Key points:
• Inform the design of future models, suggesting optimal resource allocation for training.
• Help in predicting performance improvements without extensive experimentation on large models.
1. Text Classification:
Definition: The process of assigning predefined categories to text based on its content using machine learning algorithms.
Key points:
• Commonly used in spam detection, sentiment analysis, and topic categorization.
• Techniques include supervised learning with labeled datasets and feature extraction methods.
2. Natural Language Processing (NLP):
Definition: A field of artificial intelligence focused on the interaction between computers and humans through natural language.
Key points:
• Enables machines to understand, interpret, and generate human language in a valuable way.
• Applications include chatbots, translation services, and voice-activated assistants.
3. Machine Learning:
Definition: A subset of artificial intelligence that enables systems to learn from data, identify patterns, and make decisions without explicit programming.
Key points:
• Involves algorithms like decision trees, support vector machines, and neural networks for predictive modeling.
• Used in various domains, including finance, healthcare, and marketing for data-driven insights.
4. Feature Extraction:
Definition: The process of transforming raw data into a set of measurable properties (features) that can be used for modeling.
Key points:
• Essential for reducing dimensionality and improving model performance by selecting relevant information.
• Techniques include bag-of-words, TF-IDF, and word embeddings like Word2Vec.
5. Supervised Learning:
Definition: A type of machine learning where the model is trained on labeled data, learning to predict outcomes based on input features.
Key points:
• Common algorithms include linear regression, logistic regression, and neural networks.
• Widely used in applications like image classification and medical diagnosis.
6. Unsupervised Learning:
Definition: A machine learning approach where models are trained on unlabeled data to identify patterns or groupings without predefined categories.
Key points:
• Techniques include clustering algorithms like K-means and hierarchical clustering.
• Useful for exploratory data analysis and identifying hidden structures in data.
7. Evaluation Metrics:
Definition: Quantitative measures used to assess the performance of machine learning models, guiding improvements and comparisons.
Key points:
• Common metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).
• Essential for understanding model effectiveness and making informed decisions.
8. Overfitting:
Definition: A modeling error that occurs when a machine learning model learns noise in the training data rather than the underlying pattern.
Key points:
• Results in poor generalization to new, unseen data, leading to high training accuracy but low validation accuracy.
• Techniques to prevent overfitting include cross-validation, regularization, and pruning.
9. Natural Language Generation (NLG):
Definition: The process of using algorithms to generate coherent and contextually relevant text from structured data or prompts.
Key points:
• Applications include automated report generation, content creation, and conversational agents.
• Models like GPT-3 leverage large datasets to produce human-like text.
10. Transformer Architecture:
Definition: A deep learning model architecture that uses self-attention mechanisms to process sequential data, particularly in NLP.
Key points:
• Enables parallel processing of data, improving training efficiency and performance on tasks like translation.
• Forms the basis for models like BERT and GPT, revolutionizing NLP capabilities.
11. Attention Mechanism:
Definition: A technique in neural networks that allows models to focus on specific parts of the input when making predictions.
Key points:
• Improves performance on tasks requiring context, such as machine translation and text summarization.
• Variants include self-attention and multi-head attention, enhancing model expressiveness.
12. Bag-of-Words Model:
Definition: A simplified representation of text data that disregards grammar and word order, focusing solely on word frequency.
Key points:
• Commonly used for document classification and clustering tasks.
• Limited by its inability to capture semantic meaning and context.
13. Cross-Validation:
Definition: A model evaluation technique that partitions data into subsets to test the model's performance on unseen data.
Key points:
• Helps in assessing how the results of a statistical analysis will generalize to an independent dataset.
• Common methods include k-fold and leave-one-out cross-validation.
14. Word Embeddings:
Definition: Vector representations of words that capture semantic meaning and relationships based on context and usage.
Key points:
• Techniques like Word2Vec and GloVe transform words into continuous vector spaces, enabling better model understanding.
• Essential for enhancing performance in NLP tasks by providing rich word representations.
15. Generative Models:
Definition: Models that learn to generate new data instances that resemble a training dataset, such as text or images.
Key points:
• Applications include text completion, image synthesis, and music generation.
• Popular examples include GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders).