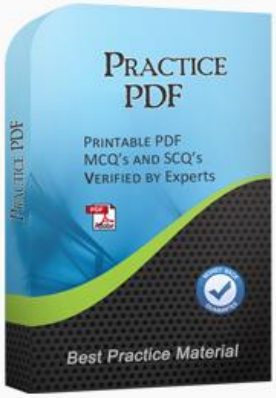
ActualtestsQuiz

ActualtestsQuiz

- Online Tool, Convenient, easy to study.
- Instant Online Access
- Supports All Web Browsers
- Practice Online Anytime
- Test History and Performance Review
- Supports Windows / Mac / Android / iOS, etc.

- Installable Software Application
- Simulates Real Exam Environment
- Builds Exam Confidence
- Supports MS Operating System
- Two Modes For Practice
- Practice Offline Anytime



- Printable PDF Format
- Prepared by IT Experts
- Instant Access to Download
- Study Anywhere, Anytime
- 365 Days Free Updates
- Free PDF Demo Available

Security & Privacy

We respect customer privacy. We use McAfee's security service to provide you with utmost security for your personal information & peace of mind.

365 Days Free Updates

Free update is available within 365 days after your purchase. After 365 days, you will get 50% discounts for updating.

Money Back Guarantee

Full refund if you fail the corresponding exam in 90 days after purchasing. And Free get any another product.

Instant Download

After Payment, our system will send you the products you purchase in mailbox in a minute after payment. If not received within 2 hours, please contact us.

Exam : **NCA-GENM**

Title : NVIDIA Generative AI Multimodal

Vendor : NVIDIA

Version : DEMO

NO.1 You are working with a large multimodal dataset containing images and text. You want to efficiently load and preprocess this data for training a generative A1 model on an NVIDIA GPU. Which of the following approaches would be most effective for maximizing data loading speed and GPU utilization?

A. Loading the entire dataset into CPU memory before starting training.

B. Using a Python-based data loader that reads images and text directly from disk during training.

C. Employing NVIDIA's DALI (Data Loading Library) to perform data loading and preprocessing on the GPU.

D. Storing the images and text in a relational database and querying the database during training.

E. Compressing the dataset into a single large archive file and decompressing it on the fly during training.

Answer: C

Explanation:

NVIDIA DALI is specifically designed for accelerating data loading and preprocessing on NVIDIA GPUs. It allows you to perform tasks like image decoding, resizing, and data augmentation directly on the GPIJ, minimizing CPIJ overhead and maximizing GPU utilizatiom Loading the entire dataset into CPU memory is impractical for large datasets. Python-based data loaders can be slow due to the GIL (Global Interpreter Lock). Querying a relational database adds overhead. Compressing the dataset can save storage space but may introduce decompression bottlenecks during training.

NO.2 Consider a scenario where you are developing a multimodal system for generating 3D models from text descriptions. The system uses a Variational Autoencoder (VAE) to generate the 3D models. During training, you observe that the generated 3D models lack diversity and tend to cluster around a few common shapes. Which of the following techniques could you employ to improve the diversity of the generated 3D models?

A. Decreasing the capacity of the VAE's latent space.

B. Increasing the weight of the Kullback-Leibler (KL) divergence term in the VAE's loss function.

C. Using a larger training dataset with more diverse text descriptions.

D. Applying techniques like adversarial training to encourage the VAE to generate more realistic 3D models.

E. Decreasing the batch size during training.

Answer: C,D

Explanation:

A larger, more diverse dataset provides more examples for the VAE to learn from, leading to more diverse generated outputs. Adversarial training can help the VAE generate more realistic and diverse 3D models by penalizing outputs that are easily distinguishable from real data. Decreasing the latent space capacity can limit the model's ability to capture the diversity of the data. Increasing the KL divergence weight can lead to underfitting and less diverse outputs. Decreasing batch size can increase the variance of gradients during training, but its impact on diversity is less direct.

NO.3 Which of the following is the MOST important factor in ensuring the 'trustworthiness' of a multimodal Generative AI model used for a safety-critical application (e.g., medical diagnosis)?

A. High accuracy on the training dataset.

B. Explainability and interpretability of the model's decisions.

C. Low computational cost for inference.

D. Use of the latest deep learning architecture.

E. Ability to generate diverse outputs.

Answer: B

Explanation:

For safety-critical applications, understanding why a model makes a certain decision is crucial. Explainability allows users to verify the model's reasoning and identify potential biases or errors. High accuracy alone is not sufficient if the model's decision-making process is opaque. While computational cost and architecture are important, they are secondary to trustworthiness in this context.

NO.4 Consider a scenario where you're building a multimodal model to generate image captions. You've pre-trained a large language model (LLM) on a massive text corpus and a convolutional neural network (CNN) on ImageNet. How would you effectively combine these pre- trained components for your image captioning task, considering the need to maintain high caption quality and training efficiency?

A. Fine-tune both the CNN and the LLM jointly on the image captioning dataset.

B. Freeze the CNN, extract image features, and train the LLM to generate captions from these features.

C. Freeze the LLM, train the CNN to predict text embeddings, and then decode these embeddings into captions.

D. Use a transformer-based encoder to process both image features and text embeddings before feeding them to the LLM decoder.

E. Train the CNN and LLM separately on unrelated datasets and then combine them at inference time using a simple averaging of their outputs.

Answer: A,D

Explanation:

Fine-tuning both the CNN and LLM jointly allows the model to adapt both visual feature extraction and language generation to the specific task of image captioning, leading to potentially higher quality captions. However, this can be computationally expensive. Using a transformer-based encoder to process both modalities before the LLM decoder allows for effective cross-modal attention and fusion, which is also a strong approach. Freezing either the CNN or LLM limits the model's ability to adapt. Training separately and averaging outputs is unlikely to produce coherent captions.

NO.5 You are developing a system that generates 3D models from text descriptions. The system currently produces models that are geometrically accurate but lack fine-grained surface details and realistic textures. Which of the following steps would be MOST effective in improving the visual realism of the generated 3D models?

A. Increase the number of polygons used to represent the 3D models.

B. Train a separate texture generation model conditioned on the text description and the generated 3D geometry.

C. Use a simpler text encoder to focus on geometric information.

D. Reduce the size of the training dataset.

E. Rely solely on procedural generation techniques.

Answer: B
Explanation:
Training a separate texture generation model allows for specializing in generating realistic surface details and textures based on both the text description and the underlying 3D geometry. Increasing polygon count (A) can help, but doesn't address texturing. Simplifying the text encoder or reducing the dataset is counterproductive. Solely relying on procedural generation might lead to lack of variability.

NO.6 You have a text-to-image model deployed using Triton Inference Server. You want to monitor the GPU utilization and inference latency to ensure optimal performance. Which of the following methods is the MOST effective way to achieve this?

A. Using 'nvidia-smi' to periodically check GPU utilization and manually calculate latency.

B. Using Triton's built-in Prometheus metrics endpoint and Grafana for visualization.

C. Using the Triton Inference Server client API to measure inference latency from the client-side.

D. Writing custom scripts to parse Triton's log files and extract performance metrics.

E. Relying solely on the operating system's resource monitor to track GPU usage.

Answer: B
Explanation:
Triton Inference Server exposes a Prometheus metrics endpoint that provides detailed information about GPIJ utilization, inference latency, and other performance metrics. Prometheus is a popular time-series database and monitoring solution. Grafana can then be used to visualize these metrics in real-time dashboards. This is the recommended approach for monitoring Triton deployments.

NO.7 You're designing a U-Net architecture for generating high-resolution medical images from low-resolution scans. Which of the following considerations are MOST crucial for maintaining fine-grained detail during the upsampling process, and how might NVIDIA's NeMo framework assist?

A. Using only bilinear interpolation in the upsampling layers to avoid introducing artifacts. NeMo can assist by providing pre-trained interpolation layers.

B. Incorporating skip connections from the contracting path to the expanding path, allowing the network to leverage high-resolution features from earlier layers. NeMo provides modules for efficient skip connection implementation and management of feature map sizes.

C. Employing a very deep network architecture to capture complex relationships between pixels. NeMo aids in managing the complexity and training of such deep networks with optimized optimizers and distributed training capabilities.

D. Using only transpose convolutional layers for upsampling to learn the optimal upsampling filters. NeMo offers optimized transpose convolution implementations for performance.

E. Ignoring the low resolution features and concentrate on better latent space sampling. NeMo can provide models to enhance sampling techniques.

Answer: B
Explanation:
Skip connections are essential in U-Nets for preserving fine-grained detail. They allow the network to access high-resolution features learned in the contracting path during the upsampling process. NeMo's features for managing skip connections and feature map sizes can streamline the implementation. While transpose convolutions (D) can be useful, they are not the most crucial without skip connections. Bilinear interpolation alone is generally insufficient for high-resolution

image generation. NeMo can aid with (C) but it's not as crucial as skip connections. (E) is incorrect because it is crucial to leverage information extracted during the downsampling process.

NO.8 You're training a multimodal model for image and text retrieval. Given an image, the model should retrieve the most relevant text description from a database, and vice-vers a. You're using a dual-encoder architecture, where one encoder processes images and the other processes text, projecting them into a shared embedding space. What is the most effective way to train the model to ensure that semantically similar images and texts have close embeddings, while dissimilar ones have distant embeddings?

A. Train the encoders independently using separate supervised tasks for image and text classification.

B. Use a contrastive loss function that minimizes the distance between embeddings of matching image-text pairs and maximizes the distance between embeddings of non-matching pairs. Example: Triplet Loss, InfoNCE.

C. Use a reconstruction loss that forces the model to reconstruct the input image from its text embedding and vice-versa.

D. Apply adversarial training to make the embeddings indistinguishable between the two modalities.

E. Use a simple L1 loss between the image and text embeddings-

Answer: B

Explanation:

Contrastive loss functions are specifically designed for learning embeddings where similarity is defined by distance. They directly encourage similar items to be close and dissimilar items to be far apart. Independent training doesn't enforce the multimodal relationship. Reconstruction loss focuses on regenerating the input, not similarity. Adversarial training aims for indistinguishability, not meaningful embeddings. L1 Loss is a basic distance metric but less effective than contrastive losses for learning semantic similarity

NO.9 Consider the following code snippet using NVIDIA Triton Inference Server. What is the purpose of the 'sequence_batching' configuration?

A. It enables batching of independent requests to improve throughput.

B. It allows for processing sequences of inputs (e.g., time series data) by maintaining state between requests.

C. It automatically scales the number of model instances based on the input load.

D. It optimizes the model for specific hardware architectures.

E. It enables dynamic batching based on request arrival times.

Answer: B

Explanation:

The 'sequence_batching' configuration in Triton Inference Server is designed to handle sequential data where the server needs to maintain state between requests. This is essential for tasks like time-series prediction or conversational AI where the context of previous inputs matters. A, E describe standard batching. C describes autoscaling and D describes model optimization.

NO.10 You are building a multi-modal model that combines text and image data for a search application. The goal is to retrieve relevant images given a text query. You have encoded both images and text into embeddings. What's a suitable loss function for training the model to ensure images

relevant to a text query are ranked higher than irrelevant ones?

A. Cross-entropy loss

B. Mean Squared Error (MSE)

C. Contrastive Loss

D. Triplet Loss

E. KL Divergence

Answer: D

Explanation:

Triplet Loss is specifically designed for ranking tasks. It takes three inputs: an anchor (text query), a positive example (relevant image), and a negative example (irrelevant image). The loss function aims to minimize the distance between the anchor and the positive example while maximizing the distance between the anchor and the negative example. Contrastive loss works with pairs, not relative rankings. Cross-entropy, MSE, and KL Divergence are not suitable for ranking problems.

NO.11 You are building a multimodal Generative A1 system to generate image captions based on both the visual content of an image and a short audio description of the scene. Which architectural approach would be MOST effective for fusing these two modalities into a coherent representation for caption generation?

A. Early Fusion: Concatenate the raw image pixel data with the raw audio waveform data before feeding it into a single model.

B. Late Fusion: Train separate image and audio encoders, then concatenate their high-level feature vectors before feeding into a caption generation model.

C. Intermediate Fusion: Train separate image and audio encoders, then use cross-attention mechanisms to allow the image features to attend to the audio features (and vice-versa) at multiple layers of the model.

D. Ignore the audio entirely, as images are sufficient for generating captions.

E. Concatenate the image file name with the audio file name before feeding into the LLM.

Answer: C

Explanation:

Intermediate Fusion, particularly using cross-attention, allows for nuanced interaction between the modalities at multiple levels of abstraction. Early fusion is generally ineffective due to the vast differences in data type. Late fusion may miss important correlations. Ignoring a modality is obviously suboptimal when aiming for multimodal understanding.

NO.12 You're building a system that takes a medical image (e.g., X-ray) and a patient's medical history (text) as input, predicting the likelihood of a specific disease. You want to use SHAP (SHapley Additive exPlanations) values to explain the model's predictions. How would you adapt SHAP to handle both image and text inputs effectively?

A. Apply KernelSHAP separately to the image and text, then combine the results.

B. Use DeepExplainer for the image component and a simple linear SHAP explainer for the text.

C. Represent both the image and text as numerical vectors and then apply a standard SHAP explainer.

D. Treat the image and text as separate models and explain each independently.

E. Use a multimodal SHAP implementation that is designed to handle both image and text features

simultaneously, considering their interaction.

Answer: E

Explanation:

The best approach is to use a multimodal SHAP implementation that considers the interaction between image and text features. This ensures a holistic explanation of the model's prediction based on both modalities. Treating them separately or simply concatenating features ignores potential synergistic effects.

NO.13 You're building a system that uses a pre-trained large language model (LLM) for generating creative stories. After deploying the system, you notice that the generated stories often contain biases present in the training data of the LLM. What are the MOST effective strategies to mitigate these biases in your generated stories? (Select TWO)

A. Increase the temperature parameter in the LLM's decoding strategy.

B. Apply bias detection and mitigation techniques to the LLM's output.

C. Fine-tune the LLM on a diverse and representative dataset.

D. Reduce the size of the LLM to minimize memory usage.

E. Use prompt engineering to steer the LLM away from biased outputs.

Answer: B,E

Explanation:

Bias detection and mitigation techniques can be applied to the LLM's output to identify and remove or modify biased content. Prompt engineering involves carefully crafting prompts to guide the LLM towards generating more balanced and unbiased stories. Fine-tuning on a diverse dataset is beneficial in the long run but is a more resource-intensive approach and doesn't guarantee the complete elimination of biases. Increasing the temperature parameter might lead to more creative but also more unpredictable and potentially biased outputs. Reducing the size of the LLM doesn't address the bias problem and might even worsen it by limiting the model's ability to capture nuanced relationships.

NO.14 You are experimenting with a multimodal model that takes both text and audio as input. During evaluation, you notice that the model is heavily biased towards the text input, largely ignoring the audio. Which of the following techniques could you employ to mitigate this modality imbalance and encourage the model to effectively utilize both inputs? (Select all that apply)

A. Increase the learning rate for the audio encoder.

B. Apply modality-specific dropout to the text encoder.

C. Use a contrastive loss function that encourages alignment between text and audio representations.

D. Reduce the size of the text encoder.

E. Replace audio features with raw audio waveform.

Answer: B,C

Explanation:

Modality imbalance is a common issue in multimodal learning. Applying modality-specific dropout to the dominant modality (text, in this case) forces the model to rely more on the other modality (audio). A contrastive loss directly encourages the model to learn aligned representations between the two modalities. Increasing the audio encoder's learning rate (A) might help, but it is less targeted than dropout or contrastive loss. Reducing the text encoder size (D) is unlikely to be helpful in a

controlled way. Replacing Audio features with raw waveform might introduce noise.

NO.15 You are analyzing the latent space of a GAN trained to generate images of human faces. You notice that interpolating between two points in the latent space often results in unrealistic or distorted faces. Which of the following techniques could potentially improve the smoothness and interpretability of the latent space?

A. Increasing the number of layers in the discriminator network.

B. Using spectral normalization in the discriminator network.

C. Applying a regularization term to the latent space during training to encourage smoothness (e.g., a Laplacian prior).

D. Decreasing the learning rate of the generator network.

E. Using a smaller batch size during training.

Answer: C

Explanation:

Regularizing the latent space directly encourages smoothness, making interpolations more realistic. Spectral normalization in the discriminator improves training stability but doesn't directly address latent space smoothness. Increasing discriminator layers or decreasing generator learning rate might influence performance, but regularization is the most direct approach. Batch size is less impactful on latent space interpretability.

NO.16 You're building a multimodal model that takes images and text as input. You notice that your model is heavily biased towards the text modality, essentially ignoring the visual input. Which of the following strategies could you employ to address this modality imbalance? (Select TWO)

A. Use a modality-specific loss function, weighting the loss from the visual modality more heavily.

B. Remove the text modality entirely.

C. Implement a gating mechanism that dynamically adjusts the contribution of each modality based on the input.

D. Increase the learning rate for the text encoder.

E. Reduce the size of the visual encoder.

Answer: A,C

NO.17 You're developing a text-to-image generation system using a pre-trained CLIP model and a diffusion model. You notice that while the generated images match the overall theme of the text prompt, they often fail to accurately represent specific objects mentioned in the prompt. What are the two MOST effective strategies to improve object fidelity in this scenario?

A. Fine-tune the diffusion model using a dataset of images specifically depicting the objects that are frequently misrepresented.

B. Increase the guidance scale during diffusion sampling, forcing the generated images to align more closely with the CLIP embeddings.

C. Replace the CLIP model with a larger, more powerful text encoder that has been trained on a more diverse dataset.

D. Implement a technique called 'Classifier-Free Diffusion Guidance', which allows for more flexible control over the generated image content.

E. All of the Above

Answer: B,D
Explanation:
Increasing the guidance scale (B) forces stronger alignment with the CLIP embeddings, improving object fidelity. Classifier-Free Diffusion Guidance (D) provides finer-grained control over image content, allowing the model to better represent specific objects. Fine-tuning the diffusion model (A) can be helpful but requires a significant amount of data. Using a larger text encoder (C) may improve overall performance but may not directly address object fidelity. Classifier-Free Diffusion Guidance and increasing guidance scale are the most targeted strategies to increase object fidelity for text-to-image models, as guidance scale can also have some artifacts.

NO.18 You are tasked with optimizing a large multimodal AI model for deployment on edge devices with limited computational resources. Which combination of techniques would provide the BEST trade-off between model accuracy and inference speed? (Select TWO)
A. Model quantization (e.g., INT8) to reduce model size and improve inference speed.
B. Increasing the number of attention heads in the transformer architecture.
C. Pruning to remove less important connections in the model.
D. Using larger batch sizes during inference to maximize GPIJ utilization.
E. Adding more layers to the model to increase its representational capacity.
Answer: A,C
Explanation:
Model quantization reduces the model size and accelerates inference by using lower-precision arithmetic. Pruning reduces the number of parameters, leading to faster computation and lower memory footprint. Increasing attention heads and adding layers increase computational cost. Larger batch sizes can improve GPU utilization on servers, but might not be feasible on resource-constrained edge devices.

NO.19 You are working with a pre-trained multimodal model that takes images and text as input. You want to fine-tune this model for a specific downstream task, but you have limited computational resources. Which of the following techniques would be most effective for reducing the memory footprint and computational cost during fine-tuning?
A. Fine-tuning the entire model with a small learning rate.
B. Freezing all layers of the pre-trained model and training only a small classification head.
C. Using quantization to reduce the precision of the model's weights and activations.
D. Applying knowledge distillation, where a smaller student model is trained to mimic the behavior of the pre-trained model.
E. Increasing the batch size to utilize the available memory more efficiently.
Answer: C,D
Explanation:
Quantization reduces the memory footprint of the model by using lower-precision representations for weights and activations. Knowledge distillation allows you to train a smaller, more efficient model that performs similarly to the larger pre-trained model. Freezing layers reduces the number of trainable parameters but may limit the model's ability to adapt to the new task. Fine-tuning the entire model, even with a small learning rate, is computationally expensive. Increasing batch size might lead to Out of Memory errors.

NO.20 You are tasked with creating a multimodal AI application that analyzes social media posts containing text, images, and user profile information to predict the likelihood of a post going viral. Which feature engineering techniques are most effective for representing and integrating these different modalities?

A. Using TF-IDF for text, pixel values for images, and one-hot encoding for user profile information.

B. Using word embeddings (e.g., Word2Vec, GloVe) for text, pre-trained CNN features (e.g., from ResNet, Inception) for images, and embedding user profiles using a graph embedding technique.

C. Using bag-of-words for text, histogram of oriented gradients (HOG) for images, and simple numerical features (e.g., number of followers) for user profiles.

D. Using character-level n-grams for text, edge detection for images, and boole an features for user profile information.

E. Using a combination of TF-IDF for text, pixel values for images, and numerical features for user profile information. Then apply PCA for dimensionality reduction.

Answer: B

Explanation:

Using word embeddings captures semantic meaning in text. Pre-trained CNN features provide high-level image representations. Graph embedding for user profiles captures relationships between users. These advanced techniques provide better representations than simple methods like TF-IDF, pixel values, or bag-of-words.

NO.21 You are building an A1 model that takes video and corresponding subtitles as input to generate short summaries of video content. Which of the following strategies are most important to reduce the chance of your model generating biased summaries? (Select all that apply)

A. Use a pre-trained language model that has been debiased.

B. Ensure the training dataset contains diverse representation of all demographic groups and viewpoints.

C. Evaluate the model's summaries on different demographic groups to identify and mitigate any disparities in performance.

D. Randomly shuffle data during training.

E. Increase the number of training epochs.

Answer: A,B,C

Explanation:

Debiasing pre-trained models helps remove existing biases. A diverse training dataset is important to reduce the influence of any single biased viewpoint. Evaluating model performance on different demographic groups allows you to find and rectify performance disparities. Random data shuffling (D) and increasing training epochs (E) do not directly address bias. Note this is a very tough question as all choices seem viable but only options, A, B and C are the correct choice.

NO.22 Which of the following techniques is MOST suitable for aligning the feature spaces of text and images in a multimodal model?

A. Using separate loss functions for text and image encoders.

B. Concatenating the features from the text and image encoders without any further processing.

C. Employing a contrastive loss function that encourages similar representations for semantically related text and images.

D. Training the text and image encoders independently.
E. Only using image data during the training process.
Answer: C
Explanation:
Contrastive loss functions are designed to bring together the representations of similar data points (e.g., a picture and its caption) while pushing apart representations of dissimilar data points. This effectively aligns the feature spaces.

NO.23 You're training a large language model (LLM) and notice that it struggles to maintain consistency and context over long passages of text. Which of the following architectural modifications would be most effective in addressing this issue?
A. Reducing the number of layers in the transformer architecture.
B. Increasing the size of the vocabulary.
C. Implementing a sparse attention mechanism to reduce computational cost
D. Increasing the maximum sequence length the model can process.
E. Using a smaller embedding dimension.
Answer: D
Explanation:
Increasing the maximum sequence length allows the model to consider a larger context window when generating text, improving its ability to maintain consistency and coherence over longer passages. Other options might address other issues, but increasing sequence length directly tackles the problem of limited context.

NO.24 You are building a multimodal model that takes images and text descriptions as input to generate new images. You want to evaluate the impact of different image encoders (ResNet50, Efficient Net) on the generated image quality and relevance to the text prompt. Which evaluation metric(s) would be MOST appropriate for this task?
A. Inception Score (IS) only
B. Frechet Inception Distance (FID) only
C. CLIP Score only
D. Both FID and CLIP Score
E. Perplexity and BLEU score
Answer: D
Explanation:
FID measures the distance between the feature distributions of generated and real images, indicating image quality and diversity. CLIP Score measures the similarity between the generated image and the text prompt, evaluating relevance. IS more suitable for evaluating unimodal image generation. Perplexity and BLEU score are for text generation.

NO.25 You are working with a multimodal dataset containing images and corresponding text descriptions. You want to train a model to generate text descriptions for new images. You decide to use a transformer-based architecture with separate encoders for images and text. How should you effectively fuse the image and text representations to enable cross-modal interaction?
A. Concatenate the final hidden states of the image and text encoders and feed them into a decoder.

B. Average the final hidden states of the image and text encoders and feed the result into a decoder.
C. Use a cross-attention mechanism where the text decoder attends to the image encoder's hidden states and vice-versa.
D. Train the image and text encoders separately and then combine their outputs using a linear layer.
E. Multiply the final hidden states of the image and text encoders and feed them into a decoder.
Answer: C
Explanation:
Cross-attention allows the decoder to selectively attend to relevant parts of both the image and text representations, enabling fine- grained interaction between the modalities. Concatenation or averaging simply combines the representations without allowing for selective attention. Training the encoders separately and then combining their outputs doesn't allow for cross modal interaction during training. Multiply operation is not standard and is not efficient.

NO.26 You have a multimodal model combining video and text data for action recognition. The model performs well on standard datasets but struggles with videos containing unusual camera angles or lighting conditions. Which data augmentation strategy would be MOST effective in improving the model's robustness?
A. Adding random noise to the audio track.
B. Randomly cropping and scaling the video frames.
C. Applying random rotations, flips, and color jittering to the video frames.
D. Replacing random words in the text descriptions with synonyms.
E. Reducing the frame rate of the videos.
Answer: C
Explanation:
Applying random rotations, flips, and color jittering directly addresses the model's sensitivity to camera angles and lighting conditions by exposing it to a wider range of visual variations during training. The other options are less relevant to these specific issues. Audio noise is irrelevant to visual robustness. Cropping and scaling are basic augmentations but less effective than transformations that simulate camera angles and lighting. Synonym replacement improves text understanding, but not visual robustness. Reducing frame rate can reduce computation but doesn't improve robustness to visual variations.

NO.27 You're building a generative A1 model that can create realistic 3D models from text descriptions. You have a dataset of text descriptions and corresponding 3D models, but the alignment between the text and the 3D models is weak. The model sometimes generates 3D shapes that don't accurately reflect the text. Which of the following techniques could improve the alignment between the text descriptions and the generated 3D models?
A. Using a contrastive loss function that encourages the model to generate 3D models that are semantically similar to the corresponding text descriptions.
B. Increasing the number of vertices and faces in the 3D models.
C. Using a pre-trained text encoder (e.g., BERT or CLIP) to extract meaningful features from the text descriptions.
D. Applying data augmentation techniques to the 3D models (e.g., random rotations and scaling).
E. Training the model with a larger batch size.

Answer: A,C
Explanation:
A contrastive loss function directly encourages the model to learn a mapping between text and 3D models that preserves semantic similarity. Using a pre-trained text encoder allows the model to leverage existing knowledge about language and extract more meaningful features from the text descriptions, improving alignment. Increasing the number of vertices and faces can improve the resolution of the models but won't directly address alignment. 3D data augmentation can improve robustness, but it's less direct. Batch size has a smaller impact compared to the other options.

NO.28 You are building a retrieval-augmented generation (RAG) system that utilizes a knowledge graph to enhance the responses generated by a large language model. The knowledge graph contains information about entities and their relationships extracted from both text documents and image metadat a. However, you observe that the system often retrieves irrelevant or outdated information from the knowledge graph, leading to inaccurate or misleading responses. Which of the following strategies would be MOST effective in addressing this issue?

A. Increase the size of the knowledge graph.

B. Implement a mechanism to filter and rank the retrieved information based on relevance and recency, using both semantic similarity and temporal information.

C. Use a simpler language model for the generative component of the RAG pipeline.

D. Reduce the number of entities in the knowledge graph.

E. None of the above.

Answer: B
Explanation:
Filtering and ranking the retrieved information based on relevance and recency ensures that the system prioritizes the most accurate and up-to-date information from the knowledge graph. Simply increasing the size of the knowledge graph or using a simpler language model would not directly address the issue of irrelevant or outdated information.

NO.29 You are fine-tuning a pre-trained language model for a specific task. You notice that the model performs well on the training data but poorly on the validation dat a. Which of the following techniques can help mitigate this overfitting problem? (Select TWO)

A. Increase the learning rate.

B. Apply weight decay (L2 regularization).

C. Use dropout regularization.

D. Decrease the batch size.

E. Increase the size of the training data.

Answer: B,C
Explanation:
Overfitting occurs when a model learns the training data too well and fails to generalize to new data. Weight decay (L2 regularization) penalizes large weights, preventing the model from becoming too complex. Dropout randomly deactivates neurons during training, forcing the model to learn more robust features. Increasing the learning rate might worsen overfitting. Decreasing the batch size can sometimes act as a regularizer, but its primary effect is on the training dynamics. While more training data is generally beneficial, if the new data is very similar to the existing training data, it won't necessarily solve the overfitting issue.

NO.30 You are tasked with evaluating a text-to-video generation model. Which of the following metrics would be MOST appropriate for assessing the temporal coherence and smoothness of the generated videos?

A. Inception Score (IS)

B. Frchet Inception Distance (FID)

C. Learned Perceptual Image Patch Similarity (LPIPS)

D. Frchet Video Distance (FVD)

E. BLEU score

Answer: D

Explanation:

Frchet Video Distance (FVD) is a metric specifically designed for evaluating video generation models. It extends the concept of FID to the video domain by comparing the distributions of features extracted from real and generated videos, taking into account the temporal dimension. IS and FID are primarily used for image generation. LPIPS measures the perceptual similarity between two images, and BLEU score is used for evaluating text generation.

NO.31 You are working with a multimodal model that combines text and video data for action recognition. The text data consists of descriptions of the actions, and the video data consists of sequences of frames. You want to fuse these modalities at a late fusion stage. Which of the following approaches BEST describes late fusion?

A. Concatenating the raw pixel values of video frames with the word embeddings of the text descriptions.

B. Training separate models for text and video data and averaging their predictions.

C. Training a single model with both text and video data as input and using a shared embedding space.

D. Training separate models for text and video data and concatenating their learned feature representations before feeding them into a final classifier.

E. Applying attention mechanisms to weigh different parts of the text and video data before feeding them into a shared model.

Answer: D

Explanation:

Late fusion involves processing each modality separately to obtain feature representations and then combining these representations at a later stage, typically by concatenation or averaging, before making a final prediction. Averaging predictions (option B) is a specific type of late fusion. Concatenating raw pixel values and word embeddings (option A) is an example of early fusion. Training a single model with a shared embedding space (option C) is also closer to early or intermediate fusion. Attention mechanisms can be used in various fusion strategies but do not define late fusion specifically.

NO.32 You are training a multimodal model with text and audio inputs. You notice that the audio modality dominates the training process, and the text modality is not contributing significantly to the final performance. Which of the following strategies can you use to address this modality imbalance? (Select TWO)

A. Increase the learning rate for the text encoder.

B. Decrease the batch size for the audio data

C. Apply a modality-specific weighting scheme to the loss function, giving more weight to the text loss

D. Remove the audio modality altogether to force the model to rely on text

E. Increase the size of the audio dataset.

Answer: A,C

Explanation:

Increasing the learning rate for the text encoder can help the text modality learn more effectively Applying a modality-specific weighting scheme to the loss function allows you to explicitly control the contribution of each modality to the overall loss, giving more weight to the underperforming text modality. Decreasing the batch size for audio data might have a small impact, but it's not a primary strategy for addressing modality imbalance. Removing the audio modality is not a desirable solution, as it eliminates valuable information. Increasing the size of audio dataset will even more dominate_ So, the most effective strategies are increasing the learning rate for text and weighting the loss function.

NO.33 You are building a Generative A1 model that generates captions for images. You want to evaluate the quality of the generated captions.

Which evaluation metrics are MOST suitable for this task?

A. Accuracy and Precision.

B. BLEU, ROUGE, and CIDEr.

C. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

D. FI-score and AUC.

E. Cosine Similarity and Euclidean Distance.

Answer: B

Explanation:

BLEU, ROUGE, and CIDEr are standard metrics used for evaluating the quality of generated text, particularly in image captioning and machine translation. These metrics compare the generated captions to reference captions and measure the similarity in terms of n-grams, word overlap, and other features. Other options are used for Classification problems (Accuracy Precision, FI-score, AUC) and Regression Problems (MSE, RMSE).

NO.34 You are building a multimodal model that takes video and audio as input. You want to fuse the information extracted from both modalities. Which of the following fusion techniques allows for learning temporal dependencies between modalities?

A. Early Fusion (concatenating features before feeding into a single network).

B. Late Fusion (averaging the probabilities from separate networks).

C. Attention-based Fusion using Transformers, allowing the model to weigh the importance of different parts of each modality over time.

D. Simple Addition of feature vectors from video and audio streams.

E. Maximum pooling across feature vectors from video and audio streams.

Answer: C

Explanation:

Attention-based Fusion, particularly using Transformers, is well-suited for capturing temporal

dependencies in multimodal data. Transformers can learn which parts of each modality are most relevant at different points in time, enabling a more nuanced fusion of information. Early Fusion (A) fuses features statically and doesn't capture temporal dependencies directly. Late Fusion (B) also struggles to capture fine- grained temporal relationships. Simple addition (D) and max pooling (E) are too simplistic to model complex temporal interactions.

NO.35 You have developed a multimodal model that predicts stock prices using news articles (text), historical stock data (time-series), and company financial reports (tabular data). You want to deploy this model using NVIDIA Triton Inference Server. Assume you have preprocessed the data and have individual models for each modality. What is the recommended approach to configure Triton for efficient and scalable multimodal inference?

A. Deploy each modality-specific model as a separate Triton model and handle the fusion logic in the client application.

B. Create a single Triton model that encapsulates the entire multimodal pipeline, including preprocessing, individual modality models, and fusion logic, using the Ensemble Modeling feature.

C. Deploy the text model using ONNX Runtime, the time-series model using TensorFlow, and the tabular data model using PyTorch, and handle fusion manually.

D. Deploy each modality-specific model as a separate Triton model and use a load balancer to distribute requests across the models.

E. Convert all models to TensorRT for maximum inference speed, even if it compromises accuracy due to quantization.

Answer: B

Explanation:

Using Triton's Ensemble Modeling feature (B) is the most efficient approach. It allows you to define a pipeline that includes preprocessing, individual modality models, and fusion logic within a single Triton model, simplifying deployment and management. This approach optimizes inter-model communication and reduces client-side overhead.