

NVIDIA.NCA-GENM.v2026-10-09.q63

Exam Code:	NCA-GENM
Exam Name:	NVIDIA Generative AI Multimodal
Certification Provider:	NVIDIA
Free Question Number:	63
Version:	v2026-10-09
# of views:	129
# of Questions views:	630
https://www.freeqas.com/qa/NVIDIA/NCA-GENM/NVIDIA.NCA-GENM.v2026-10-09.q63.html	

NEW QUESTION: 1

Consider a multimodal generative A1 model that produces images based on textual prompts. The model is prone to generating images that are similar to those in the training data, resulting in a lack of novelty. Which hyperparameter adjustment would be MOST effective in increasing the diversity of the generated images?

- A. Decrease the batch size during inference.
- B. Increase the temperature parameter in the decoding process.
- C. Reduce the learning rate during fine-tuning.
- D. Decrease the number of layers in the decoder network.
- E. Increase the weight decay during training.

Answer: B ([LEAVE A REPLY](#))

Increasing the temperature parameter during decoding introduces more randomness into the sampling process, leading to a wider range of possible outputs and thus increasing the diversity of generated images. Decreasing batch size doesn't directly affect diversity. Reducing the learning rate affects training, not the diversity of generated images after training. Decreasing layers might reduce model capacity but not necessarily increase diversity. Increasing weight decay can prevent overfitting but doesn't directly address the lack of novelty in generated images.

NEW QUESTION: 2

Which data augmentation techniques are MOST suitable for improving the robustness of a multimodal model that uses images and text?

- A. Randomly cropping images and translating text to different languages.
- B. Rotating images and back-translating text (translating to another language and back to the original).
- C. Adding Gaussian noise to images and randomly deleting words in the text.
- D. Changing the image resolution and increasing the text size.
- E. Rescaling images and changing the font of the text.

Answer: ([SHOW ANSWER](#))

Rotating images and back-translating text are effective as they introduce variations that the model might encounter in real-world scenarios. Adding Gaussian noise and randomly deleting words helps the model become more robust to noisy or incomplete data. Cropping and translation, while augmentation techniques, don't specifically target multimodal robustness as effectively. Changing resolution and font are less likely to generalize well.

NEW QUESTION: 3

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: ([SHOW ANSWER](#))

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.

NEW QUESTION: 4

You are training a text-to-image diffusion model and observe that the generated images often exhibit a 'washed-out' or overly smooth appearance. Which of the following adjustments to the training process would likely improve the image quality and detail?

- A. Decrease the number of diffusion steps used during training.
- B. Increase the weight of the perceptual loss function in the training objective.
- C. Reduce the batch size used during training to minimize memory consumption.
- D. Apply more aggressive data augmentation techniques to the training dataset.
- E. Reduce the learning rate for the U-Net architecture within the diffusion model.

Answer: B ([LEAVE A REPLY](#))

A perceptual loss function encourages the generated images to have more realistic features and details, as it compares the high-level representations of the generated images to the real images. Increasing its weight in the training objective would incentivize the model to produce more detailed and visually appealing results. Decreasing diffusion steps leads to faster but often lower-quality results. Reducing batch size can affect training stability but doesn't directly address the 'washed-out' appearance. Data augmentation and learning rate adjustments may have some impact, but are less directly targeted at improving image detail.

NEW QUESTION: 5

You're designing a generative AI system to create realistic 3D models of furniture from text descriptions. Which of the following approaches would likely yield the MOST realistic and detailed results, and how can NVIDIA's tools contribute to its success?

- A.** Using a simple variational autoencoder (VAE) trained on a dataset of 3D furniture models, without any text-based guidance. NVIDIA's GPUs can accelerate the VAE training process.
- B.** Employing a text-to-image model like Stable Diffusion to generate 2D images of the furniture from different viewpoints, and then using multi-view stereo reconstruction to create a 3D model. NVIDIA's GPUs can accelerate both the image generation and reconstruction processes.
- C.** Training a generative adversarial network (GAN) to directly generate 3D meshes from text descriptions, using a differentiable renderer as part of the discriminator. NVIDIA's GPUs are essential for training GANs with differentiable renderers.
- D.** Using a rule-based system to procedurally generate 3D furniture models based on keywords extracted from the text descriptions. NVIDIA's PhysX engine can be used to simulate realistic physics interactions.
- E.** None of the above

Answer: ([SHOW ANSWER](#))

Directly generating 3D meshes from text using a GAN with a differentiable renderer (C) allows the model to learn complex relationships between text and 3D geometry. Differentiable rendering enables the discriminator to evaluate the realism of the generated 3D models. VAEs (A) are less capable of generating high-detail models. Multi-view stereo (B) can be effective, but relies on the quality of the 2D images. Rule-based systems (D) lack the flexibility to capture the nuances of natural language. NVIDIA GPUs are crucial for the computationally intensive GAN training and differentiable rendering processes. GANs are difficult to train. The best option would be to directly train them on NVIDIA GPU and a Differentiable renderer.

NEW QUESTION: 6

You are building a generative AI model that creates realistic product designs based on textual descriptions and a reference image depicting a similar, but not identical, product. You are using a Variational Autoencoder (VAE) architecture. However, the generated images lack the fine-grained details present in the reference image. Which of the following methods would be most suitable to incorporate fine-grained details from the reference image into the generated design?

- A.** Increase the latent space dimensionality of the VAE.

- B. Use a larger convolutional kernel Size in the decoder
- C. Implement a skip connection from the encoder of the reference image to the decoder of the generative model, allowing the decoder to directly access features from the reference image at multiple scales.
- D. Reduce the batch size during training.
- E. Replace the VAE with a Generative Adversarial Network (GAN).

Answer: C (LEAVE A REPLY)

Skip connections are designed to pass information from earlier layers to later layers, preserving fine-grained details. In this multimodal scenario, skip connections from the reference image encoder allow the generator to directly access and utilize those details. Increasing latent space dimensionality might capture more information, but doesn't guarantee fine-grained detail transfer. Larger kernels can blur details. Reducing batch size helps with generalization but doesn't directly address detail transfer. GANs might generate sharper images, but lack the explicit control over detail transfer that skip connections provide.

NEW QUESTION: 7

What is a main application of Triton Inference Server?

- A. Triton Server can be used to generate images from pure noise.
- B. Triton Server can be used to deploy AI models on the GPU only.
- C. Triton Server can be used to execute GPU-accelerated graph analysis with cuGraph.
- D. Triton Server can be used to deploy neural networks from various frameworks.

Answer: D (LEAVE A REPLY)

NVIDIA Triton Inference Server is an open-source model-serving platform designed to standardize production deployment across heterogeneous model formats and hardware backends. Its defining capability is multi-framework support: a single Triton instance can concurrently serve models trained in TensorFlow, PyTorch, ONNX Runtime, TensorRT, OpenVINO, and custom Python/C++ backends, exposing them through unified HTTP/REST and gRPC inference APIs. This eliminates the need for framework-specific serving stacks and lets teams standardize deployment infrastructure independent of how a given model was trained. Option B is a common but incorrect assumption - Triton explicitly supports both GPU and CPU inference, which matters for cost-sensitive or edge deployments where GPU availability is limited. Option C confuses Triton with cuGraph, a separate RAPIDS library for GPU-accelerated graph analytics (unrelated to model serving), and option A describes a generative task (denoising diffusion), which is a *model capability*, not a Triton *server* function - Triton can host such a model, but "generating images from noise" is not what the server itself does.

Triton also provides dynamic batching, concurrent model execution, model ensembling (chaining pre/post- processing with inference), and metrics export - features tested elsewhere in the Software Development and Engineering and Performance Optimization domains.

Reference: Software Development and Engineering domain - Triton Inference Server, multi-framework model deployment.

NEW QUESTION: 8

You have developed a multimodal model that uses both audio and video data to detect human emotions. During testing, you observe that the model performs exceptionally well on controlled lab recordings but poorly in real-world scenarios with background noise and varying lighting conditions. What technique would be MOST effective in improving the model's generalization ability to real-world data?

- A. Reducing the model's complexity to prevent overfitting to the lab recordings.
- B. Increasing the amount of data from lab recordings.
- C. Data augmentation techniques such as adding noise to the audio, simulating different lighting conditions for the video, and using transfer learning from pre-trained audio and video models.
- D. Training separate models for lab recordings and real-world data.
- E. Replacing the audio input with text transcripts.

Answer: ([SHOW ANSWER](#))

Data augmentation is the most effective way to improve a model's generalization ability to real-world data. By adding noise to the audio, simulating different lighting conditions for the video, we can create a more diverse training dataset that is more representative of the real world. Also leveraging pre-trained audio and video models helps to leverage the knowledge learned on large datasets.

NEW QUESTION: 9

You are developing a multimodal system for medical diagnosis that integrates patient history (text), X-ray images, and heart rate data (time-series). A significant portion of the heart rate data is missing due to sensor failures. What is the MOST appropriate method to handle this missing data to ensure the model's accuracy and prevent bias?

- A. Replace the missing heart rate data with the mean heart rate value calculated from the available data.
- B. Remove the records of patients with missing heart rate data from the dataset.
- C. Use a time-series imputation technique, such as Kalman filtering or recurrent neural networks, to estimate the missing heart rate values based on the available data and temporal patterns.
- D. Assign a fixed, arbitrary value (e.g., 0) to all missing heart rate data points.
- E. Train a separate model specifically on data without the time-series component.

Answer: C ([LEAVE A REPLY](#))

Imputation using time-series techniques (C) is the most suitable method as it leverages the temporal dependencies within the heart rate data to estimate missing values, minimizing bias and preserving the integrity of the data. Mean imputation or arbitrary value assignment can introduce significant bias, and removing records reduces the dataset size.

NEW QUESTION: 10

You are building a multimodal model to predict stock prices using financial news articles (text), historical stock prices (time-series), and company logos (images). You have preprocessed the

data and are ready to train your model. Which of the following architectures would be MOST suitable for effectively integrating these three modalities?

- A.** A simple feed forward neural network with concatenated features from all modalities.
- B.** A model that combines a Transformer for text, an LSTM for time-series, and a CNN for images, with a late fusion strategy using a weighted averaging of predictions.
- C.** A model that uses a Transformer encoder for each modality, followed by a shared Transformer decoder for prediction, enabling cross-modal attention at the decoder level.
- D.** Separate models for each modality trained independently, and then ensembled together at the prediction stage.
- E.** A model that converts all data into a single text format and uses a large language model (LLM) for prediction.

Answer: B,C ([LEAVE A REPLY](#))

Combining a Transformer for text, an LSTM for time-series, and a CNN for images with a late fusion approach allows each modality to be processed by a suitable architecture and then combined to generate a final prediction. Using transformers in each modality with shared Transformer decoder can efficiently integrate and predict stock prices using cross modal attention . A simple feedforward network is unlikely to capture the temporal dependencies in the time-series data or the complex relationships between modalities. Ensembling independent models doesn't allow for cross-modal learning. Converting all data into text might lose valuable information from the other modalities. Therefore, hybrid architecture combining transformers, LSTMs, and CNNs with cross-modal attention or late fusion would be most effective.

NEW QUESTION: 11

Assume you have trained a text-to-image diffusion model using a large dataset of landscape photographs. You now want to adapt this model to generate images of photorealistic portraits. Which of the following fine-tuning strategies is most likely to yield the best results with the least amount of training data and time?

- A.** Retrain the entire diffusion model from scratch using the portrait dataset.
- B.** Fine-tune only the CLIP model with portrait-related text descriptions and corresponding images.
- C.** Fine-tune the IJ-Net architecture of the diffusion model while keeping the CLIP model fixed.
- D.** Fine-tune both the CLIP model and the U-Net architecture with the portrait dataset, using a smaller learning rate than the initial training.
- E.** Only fine tune the final layer of the IJ-Net model with the portrait dataset.

Answer: ([SHOW ANSWER](#))

Fine-tuning both the CLIP model and the IJ-Net architecture is the most effective approach. The CLIP model needs to learn the semantic relationship between portrait-related text and images, and the U-Net needs to adapt to generating portraits instead of landscapes. Using a smaller learning rate prevents overfitting and allows the model to leverage its existing knowledge from the landscape dataset. Retraining from scratch is wasteful, and fine-tuning only one component may not be sufficient for good performance. Simply fine-tuning the last layer will not change much.

NEW QUESTION: 12

You're training a multimodal model to generate images from text prompts. The model architecture consists of a text encoder (Transformer) and an image decoder (GAN). After training, you observe that the generated images are highly realistic but often don't accurately reflect the details specified in the text prompt. What strategy would be MOST effective in improving the alignment between the text prompts and the generated images?

- A. Increase the capacity of the GAN's generator network.
- B. Use a larger dataset of images for training the GAN.
- C. Introduce a contrastive loss that encourages the image embedding to be close to the text embedding of its corresponding prompt and far from the embeddings of other prompts.
- D. Reduce the learning rate of the text encoder.
- E. Use a simpler text encoder.

Answer: C ([LEAVE A REPLY](#))

A contrastive loss explicitly encourages the model to learn a shared embedding space where images and their corresponding text prompts are close together, while unrelated images and prompts are pushed apart. This directly addresses the alignment problem. Increasing GAN capacity or dataset size might improve image quality, but not necessarily text-image alignment. Reducing the text encoder learning rate might slow down training but doesn't guarantee better alignment. A simpler encoder will likely hurt performance.

NEW QUESTION: 13

You are tasked with building a system that generates realistic images from text descriptions. Which of the following loss functions is MOST crucial for ensuring the generated images are both visually appealing and semantically aligned with the text?

- A. Mean Squared Error (MSE) between the generated image and the target image (if available).
- B. Binary Cross-Entropy loss for discriminating between real and generated images (used in GANs).
- C. A perceptual loss that measures the difference in high-level features extracted by a pre-trained convolutional neural network.
- D. A contrastive loss that encourages similar text and image embeddings to be closer together in a shared embedding space.
- E. All of the above.

Answer: E ([LEAVE A REPLY](#))

Option E is the best answer. All the mentioned loss functions play a vital role. GANs use Binary Cross-Entropy for realistic image generation. Perceptual Loss helps in creating more visually appealing images, and Contrastive Loss ensures that the image aligns with the text description by projecting them into a common embedding space.

NEW QUESTION: 14

In machine learning, what is the purpose of data normalization?

- A. To remove irrelevant data from the dataset.
- B. To increase the complexity of the dataset.
- C. To convert data into a specific format for easier analysis.
- D. To reduce the dimensionality of the dataset.

Answer: C ([LEAVE A REPLY](#))

Normalization rescales numeric features onto a common, well-defined range or distribution - for example, min-max scaling to $[0,1]$, or standardization to zero mean and unit variance (z-score) - so that features measured on different scales contribute comparably to model training. Among the options given, "converting data into a specific format for easier analysis" is the closest description of this rescaling purpose, though the more precise technical framing is: normalization standardizes the scale of feature values to stabilize and accelerate optimization.

This matters mechanically because many algorithms are scale-sensitive: gradient descent converges faster and more stably when input features share a comparable range (large-scale features would otherwise dominate the loss gradient), distance-based methods (k-NN, k-means, SVMs with RBF kernels) require comparable scales for distance calculations to be meaningful, and regularization terms penalize weight magnitude uniformly, which only makes sense if inputs are on comparable scales.

It is important to distinguish normalization from the other three options: it does not remove data (A, which is cleansing/filtering), does not increase complexity (B, the opposite of its intent), and does not reduce dimensionality (D, which describes techniques like PCA or feature selection - an entirely separate preprocessing goal focused on the number of features, not their scale).

Reference: Core Machine Learning and AI Knowledge domain - feature scaling (normalization, standardization) vs. dimensionality reduction.

NEW QUESTION: 15

Consider the following Python code snippet using PyTorch. What does this code do in the context of data preprocessing for a Generative AI model?

```
import torch
import torchvision.transforms as transforms

transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

A.])

```
import torch
import torchvision.transforms as transforms

transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

B.

```
import torch
import torchvision.transforms as transforms

transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

C.

```
import torch
import torchvision.transforms as transforms

transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

D.

```
import torch
import torchvision.transforms as transforms

transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

E.

Answer: ([SHOW ANSWER](#))

The code snippet first resizes the images to a fixed size (256x256). Then, it converts the images into PyTorch tensors, which are the standard data format for PyTorch models. Finally, it normalizes the pixel values to a range of approximately $[-1, 1]$. This normalization helps to improve the training stability and performance of the generative A1 model by scaling the input values.

NEW QUESTION: 16

Consider a scenario where you are developing a virtual assistant that can answer questions about images. You have a large dataset of images and corresponding question-answer pairs. Which architecture is BEST suited for this task?

- A. A convolutional neural network (CNN) for image feature extraction followed by a recurrent neural network (RNN) for question encoding and a fully connected layer for answer prediction.
- B. A transformer-based model that processes both images and questions as sequences of tokens, allowing for attention-based interaction between modalities.
- C. A simple feedforward neural network that takes flattened image pixels and question embeddings as input.
- D. A combination of a pre-trained word embedding model (e.g., Word2Vec) for the question and a separate model for image classification.
- E. A Support Vector Machine (SVM) trained on image features and question keywords.

Answer: B ([LEAVE A REPLY](#))

Option B, a transformer-based model, is the most suitable architecture for Visual Question Answering (VQA). Transformers excel at capturing long-range dependencies and interactions between different modalities (image and text) using attention mechanisms, leading to better performance than CNN-RNN combinations or simpler models.

Valid NCA-GENM Dumps shared by PrepPdf.com for Helping Passing NCA-GENM Exam! PrepPdf.com now offer the **newest NCA-GENM exam dumps**, the PrepPdf.com NCA-GENM exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com NCA-GENM dumps with Test Engine here:
<https://www.preppdf.com/NVIDIA/NCA-GENM-prepaway-exam-dumps.html> (**58 Q&As Dumps**, **40%OFF Special Discount: Exam-Tests**)

NEW QUESTION: 17

You have a multimodal model that processes images and text, and you want to deploy it on an edge device with limited computational resources. Which of the following hardware acceleration strategies would be MOST effective in improving the model's inference speed on the edge device?

- A. Using NVIDIA TensorRT to optimize the model for the specific edge device.
- B. Implementing distributed inference across multiple edge devices.
- C. Using a larger batch size to improve GPU utilization.
- D. Converting the model to a smaller architecture with fewer parameters, accepting a lower accuracy.
- E. Offloading complex computations to a cloud server.

Answer: ([SHOW ANSWER](#)**)**

NVIDIA TensorRT is specifically designed to optimize models for NVIDIA GPUs, which are commonly found in edge devices. Converting to a smaller architecture reduces the computational burden on the edge device. Distributed inference is complex to setup and cloud offloading defeats the purpose of edge deployments. A larger batch size requires more memory, which can be limiting on edge devices. Accepting a lower accuracy can improve inference, which can be acceptable for certain edge deployments

NEW QUESTION: 18

You are building a system to generate captions for images. You want to evaluate how well the generated captions describe the content of the images. Which of the following metrics are most suitable for evaluating the quality of image captions?

- A. Pixel Accuracy.
- B. BLEU (Bilingual Evaluation Understudy).
- C. Inception Score.
- D. ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- E. F 1-Score

Answer: B,D ([LEAVE A REPLY](#))

BLEU and ROUGE are standard metrics for evaluating the quality of generated text, especially in the context of machine translation and text summarization. BLEU measures the precision of n-grams in the generated text compared to reference texts, while ROUGE measures the recall. Pixel accuracy and Inception Score are more relevant for image classification and image generation tasks, respectively. F1-Score could be used if you manually labeled different image aspects of the caption.

NEW QUESTION: 19

You have a dataset containing information about sales performance for different regions in the last ten years.

Which type of data visualization would be most appropriate to compare the sales performance across regions on a year-by-year basis?

- A. Scatter plot
- B. Line chart
- C. Bar chart
- D. Pie chart

Answer: ([SHOW ANSWER](#)**)**

Reviewer note: Marked answer (D, pie chart) is inconsistent with standard data-visualization practice for year-by-year, multi-region comparison; a line chart (B) is the technically defensible choice.

I need to flag this one directly: the marked answer (D, pie chart) does not hold up technically, and I won't present it as correct just because it's what the answer key says. A pie chart shows the proportional breakdown of a whole at a single point in time - it has no mechanism for representing

a trend across ten years, and using ten overlapping pie charts (one per year) to compare regional performance would be one of the least readable choices available, not the most appropriate. The technically correct choice is a line chart (B): with ten years of data per region, a line chart plots each region as a separate series across a shared time axis, making year-over-year trends, growth rates, inflection points, and cross-region divergence immediately visible - exactly the "year-by-year" comparison the question specifies. A grouped/clustered bar chart (C) is a reasonable secondary choice if the emphasis is discrete year-to-year comparison rather than continuous trend, but it becomes visually cluttered with ten years

× multiple regions. A scatter plot (A) is better suited to examining the relationship between two continuous variables (e.g., sales vs. marketing spend) than to a time-series comparison across categories.

If this exact answer appears on a live exam or official material, treat D with skepticism - this explanation reflects standard data visualization practice, not the source document's marked key.

NEW QUESTION: 20

You are building a text-to-image generation pipeline using CLIP and a diffusion model. After training, you notice that the generated images often lack the specific details mentioned in the text prompts. Which of the following strategies could you employ to improve the alignment between text and image?

- A. Increase the number of layers in the I-I-Net architecture of the diffusion model.
- B. Fine-tune the CLIP model on a dataset of text-image pairs relevant to your desired domain.
- C. Use negative prompt engineering to guide the diffusion process away from undesired attributes.
- D. Increase the number of diffusion steps during the image generation process.
- E. All of the above.

Answer: ([SHOW ANSWER](#))

All the strategies mentioned can help improve the alignment between text and image. Increasing U-Net layers can improve image detail. Fine-tuning CLIP improves semantic understanding. Negative prompts refine image generation. More diffusion steps can improve image quality. All options contribute to better alignment.

NEW QUESTION: 21

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 ([LEAVE A REPLY](#))

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.

NEW QUESTION: 22

You have trained a text-to-image diffusion model. During inference, you notice that the generated images often lack fine-grained details and appear blurry. Which of the following techniques could you apply to improve the image quality without retraining the model?

A. Increase the number of diffusion steps during sampling.

B. Decrease the guidance scale during sampling.

C. Use a larger batch size during inference.

D. Reduce the learning rate during sampling.

E. Increase the model's capacity by adding more layers.

Answer: A ([LEAVE A REPLY](#))

Increasing the number of diffusion steps during sampling allows the model to refine the generated image more thoroughly, leading to finer details and reduced blurriness. The guidance scale controls how closely the generated image adheres to the input text prompt; increasing it typically improves adherence but can sometimes reduce diversity. Batch size primarily affects computational efficiency. Reducing the learning rate is relevant during training, not inference. Adding model layers requires retraining.

NEW QUESTION: 23

Which technique is commonly used to speed up AI model training and inference on hardware accelerators?

A. Quantization

B. Data augmentation

C. Model enlargement

D. Dropout

Answer: ([SHOW ANSWER](#))

Quantization reduces the numerical precision used to represent model weights and activations - for example, converting FP32 weights to INT8 - which decreases memory bandwidth requirements and allows hardware accelerators (GPU Tensor Cores, dedicated INT8 inference engines) to execute more operations per cycle, directly speeding up both training (in its mixed-precision form) and, especially, inference. Post-training quantization and quantization-aware

training are the two dominant approaches, with the latter simulating quantization effects during training to better preserve accuracy at reduced bit-widths. NVIDIA's TensorRT relies heavily on quantization (alongside layer fusion and kernel auto-tuning) to accelerate deployed inference. The distractors describe techniques that serve entirely different purposes: data augmentation (B) increases training data diversity to improve generalization, not computational speed - it typically adds preprocessing overhead rather than reducing it. Model enlargement (C) does the opposite of speeding up computation; larger models require more FLOPS and memory, increasing latency. Dropout (D) is a regularization technique applied during training to prevent overfitting by randomly zeroing activations - it has no role in inference-time speed (and is typically disabled at inference) and does not meaningfully accelerate training compute either. Quantization is frequently paired with pruning and kernel/operator fusion as the three core techniques for hardware-accelerated performance optimization. Reference: Performance Optimization domain - quantization, TensorRT, inference acceleration techniques.

NEW QUESTION: 24

You are tasked with generating realistic images of human faces using a GAN. However, you notice that the generated images often contain artifacts, such as distorted facial features or unrealistic textures. Which of the following techniques would be most effective in improving the realism and quality of the generated faces?

- A. Using a smaller batch size.
- B. Training the GAN for fewer epochs.
- C. Employing a StyleGAN architecture with adaptive instance normalization (AdaIN) and mapping network.
- D. Using a simpler discriminator architecture.
- E. Applying L1 regularization to the generator's weights.

Answer: ([SHOW ANSWER](#))

StyleGAN architecture, with its AdaIN and mapping network, is specifically designed to control and manipulate the style attributes of generated images, leading to more realistic and high-quality outputs, particularly for complex structures like human faces. AdaIN helps in normalizing feature statistics based on style codes, enabling fine-grained control over the visual appearance.

NEW QUESTION: 25

When training a Variational Autoencoder (VAE) for generating new data points, which of the following objectives does the VAE optimize?

- A. Maximizing the likelihood of the input data given the latent representation.
- B. Minimizing the Kullback-Leibler (KL) divergence between the learned latent distribution and a prior distribution (e.g., a Gaussian distribution).
- C. Maximizing the similarity between the input data and the reconstructed data.
- D. All of the above.
- E. Only A and B.

Answer: (SHOW ANSWER)

A VAE optimizes all three objectives. It aims to maximize the likelihood of the input data given the latent representation (reconstruction accuracy), minimize the KL divergence to ensure the latent space is well-structured and smooth, and maximize the similarity between the input and reconstructed data to ensure effective encoding and decoding.

NEW QUESTION: 26

You are building a multimodal model to translate spoken language into sign language animations. You have audio recordings of spoken words and corresponding sign language video sequences. Which architecture would be most suitable for this task?

- A.** A simple feedforward neural network.
- B.** A convolutional neural network (CNN) for audio and another CNN for video, followed by a fully connected layer.
- C.** A recurrent neural network (RNN) or Transformer network with attention mechanisms to capture the temporal dependencies in both the audio and video sequences, and a connectionist temporal classification (CTC) loss function for alignment.
- D.** A generative adversarial network (GAN) to generate realistic sign language animations.
- E.** A support vector machine (SVM) to classify audio and video features.

Answer: C (LEAVE A REPLY)

RNNs or Transformer networks with attention mechanisms are best suited for sequence-to-sequence tasks like translation, as they can capture temporal dependencies. CTC loss helps align the audio and video sequences, even if they are not perfectly synchronized. CNNs are good for feature extraction but don't handle sequential data as well as RNNs/Transformers.

NEW QUESTION: 27

You are building a conditional GAN (cGAN) to generate images conditioned on text descriptions. The generator takes a noise vector and a text embedding as input. Which of the following approaches would be most effective for combining the noise vector and text embedding before feeding them into the generator's first layer?

- A.** Concatenating the noise vector and text embedding directly.
- B.** Adding the noise vector and text embedding element-wise.
- C.** Applying a learned linear transformation to both the noise vector and text embedding, then concatenating the results.
- D.** Treating the text embedding as a set of attention weights applied to the noise vector.
- E.** Using a cross-attention mechanism where the noise vector attends to the text embedding.

Answer: C (LEAVE A REPLY)

Applying learned linear transformations before concatenation allows the model to learn optimal projections of both the noise vector and text embedding into a shared space, improving the generator's ability to incorporate the conditioning information effectively. Directly concatenating can work, but may not be as effective as learned transformations. Other options are less common and less directly applicable to the task.

NEW QUESTION: 28

You are working on a multimodal model for video captioning, where the model needs to generate captions describing the actions and events happening in a video. You notice that the model tends to focus only on the most salient objects in the scene and ignores subtle but important actions. Which of the following techniques can help the model attend to these subtle actions and generate more comprehensive captions?

- A. Increasing the learning rate during training.
- B. Using a larger batch size.
- C. Implementing a hierarchical attention mechanism that first attends to relevant time steps and then to relevant regions within those time steps.
- D. Adding more layers to the LSTM or GRU used for sequence modeling.
- E. Decreasing the regularization strength.

Answer: C ([LEAVE A REPLY](#))

A hierarchical attention mechanism is the MOST appropriate technique. By first attending to relevant time steps (which might contain the subtle actions) and then attending to relevant regions within those time steps, the model can focus on the specific parts of the video that are most informative for describing the subtle actions. Increasing learning rate (A), using a larger batch size (B), adding more layers (D), and decreasing regularization strength (E) are unlikely to solve the problem of attending to subtle actions specifically. These can all improve performance, but they don't address the attention mechanism itself.

NEW QUESTION: 29

You are building a system to translate spoken language into images. You have a large dataset of audio clips and corresponding images.

Which of the following is the MOST appropriate architecture?

- A. A sequence-to-sequence model with an LSTM encoder for the audio and an LSTM decoder for generating image pixels directly.
- B. A CNN for audio feature extraction, followed by a GAN for generating images conditioned on those features.
- C. A transformer-based model that attends to both audio features and a learned visual vocabulary to generate images.
- D. A hidden Markov model (HMM) trained to map audio features to image segments.
- E. A Support Vector Machine (SVM) trained on audio features to classify the type of image to generate.

Answer: C ([LEAVE A REPLY](#))

Option C is the most appropriate. Transformer models can effectively handle sequence-to-sequence tasks and leverage attention mechanisms to capture the relationship between audio and visual features. The learned visual vocabulary helps in generating more coherent and realistic images. GANs (option B) could be part of the system but would likely need a transformer to provide the conditioned features.

NEW QUESTION: 30

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 ([LEAVE A REPLY](#))

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.

NEW QUESTION: 31

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-versa. 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 ([LEAVE A REPLY](#))

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

Valid NCA-GENM Dumps shared by PrepPdf.com for Helping Passing NCA-GENM Exam! PrepPdf.com now offer the **newest NCA-GENM exam dumps**, the PrepPdf.com NCA-GENM exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com NCA-GENM dumps with Test Engine here:

<https://www.preppdf.com/NVIDIA/NCA-GENM-prepaway-exam-dumps.html> (58 Q&As Dumps,

40%OFF Special Discount: Exam-Tests)

NEW QUESTION: 32

You are working with a multimodal model that combines text and image inputs. You want to analyze the model's attention mechanisms to understand which parts of the image are most relevant to specific words in the input text. What technique can you use to visualize and interpret the model's attention weights in this scenario?

- A. t-SNE (t-distributed Stochastic Neighbor Embedding)
- B. PCA (Principal Component Analysis)
- C. Attention Heatmaps
- D. ROC curves (Receiver Operating Characteristic curves)
- E. Confusion Matrix

Answer: C ([LEAVE A REPLY](#))

Attention heatmaps are a visualization technique used to highlight the regions of an image that the model is focusing on when processing specific words in the input text. By overlaying the attention weights onto the image, you can identify the most relevant areas. t-SNE and PCA are dimensionality reduction techniques used for visualizing high-dimensional data in lower dimensions. ROC curves and confusion matrices are used to evaluate the performance of classification models.

NEW QUESTION: 33

A research team has developed a novel multimodal model that fuses text, image, and audio data. They want to quantitatively evaluate the model's performance in comparison to several existing state-of-the-art models. Which of the following evaluation metrics would be MOST appropriate to assess the model's ability to generate coherent and relevant text descriptions based on the combined multimodal input?

- A. Perplexity.
- B. Inception Score.
- C. BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- D. Structural Similarity Index Measure (SSIM).

E. Frechet Inception Distance (FID).

Answer: ([SHOW ANSWER](#))

BLEU and ROUGE are standard metrics for evaluating text generation tasks by comparing the generated text to reference texts. They assess the similarity and overlap in terms of n-grams. Perplexity measures the uncertainty of a language model. Inception Score and FID are used for evaluating image generation quality. SSIM measures the similarity between two images.

NEW QUESTION: 34

You are developing a multimodal model that combines time-series data from sensor readings with natural language descriptions of events. The time-series data has varying sampling rates and the text descriptions are often vague and ambiguous. How would you best address the challenge of aligning and fusing these two modalities to improve model performance?

- A. Resample the time-series data to a uniform sampling rate and directly concatenate it with the text embeddings.
- B. Ignore the time-series data and train the model only on the text descriptions.
- C. Use a dynamic time warping (DTW) algorithm to align the time-series data with the text descriptions and then use a cross-modal attention mechanism for fusion.
- D. Average the time-series data over a fixed time window and concatenate it with the text embeddings.
- E. Train separate models for time-series and text and average their predictions.

Answer: C ([LEAVE A REPLY](#))

DTW helps align time-series data with varying lengths and temporal distortions to text. Cross-modal attention then effectively fuses the aligned modalities, allowing the model to learn relationships between them. Resampling and direct concatenation (A) doesn't account for temporal variations. Ignoring data (B) is counterproductive. Averaging (D) loses temporal information. Averaging separate model outputs (E) is a form of late fusion and less effective than joint learning after alignment.

NEW QUESTION: 35

You're training a multimodal model for generating stories from images and audio. You use a Transformer architecture. During training, you notice that the model struggles to maintain long-range dependencies in the generated stories, leading to incoherent narratives. Which of the following techniques would be MOST effective in addressing this issue within the Transformer architecture?

- A. Reducing the number of layers in the Transformer.
- B. Using a smaller embedding dimension.
- C. Incorporating positional encodings and increasing the attention window size.
- D. Removing the self-attention mechanism.
- E. Using only audio as input.

Answer: C ([LEAVE A REPLY](#))

Positional encodings help the Transformer understand the order of words in the sequence, which is crucial for maintaining coherence. Increasing the attention window size allows the model to attend to a larger context when generating each word, enabling it to capture longer-range dependencies. Reducing layers or embedding dimension would likely worsen the problem. Removing self-attention would defeat the purpose of using a Transformer. Positional encodings and attention window size are key to transformer performance with respect to long range dependencies.

NEW QUESTION: 36

Consider this Python code snippet using PyTorch:

A.

```
import torch
text_embedding = torch.randn(32, 256) # Batch size 32, embedding dimension 256
image_embedding = torch.randn(32, 512) # Batch size 32, embedding dimension 512

# Attempting cross-modal attention
attention = torch.matmul(text_embedding, image_embedding.transpose(0, 1))
print(attention.shape)
```

- B. torch.Size([32, 512]). The issue is a dimension mismatch.
- C. torch.Size([256, 512]). This implementation is correct and efficient for cross-modal attention.
- D. torch.Size([32, 32]). This is correct and computes attention weights for each text-image pair in the batch independently
- E. Error. The transpose operation is incorrect for achieving cross-modal attention.

Answer: A ([LEAVE A REPLY](#))

The shape of the 'attention' tensor is torch.Size([32, 32]). The matrix multiplication of (32, 256) with (512, 32) results in a (32, 32) tensor. The crucial issue here is the batch-wise attention calculation. The attention weights are being computed between all text embeddings and all image embeddings in the batch. During training, this leads to 'information leakage' because the model is learning relationships between samples that shouldn't be related (i.e., different text-image pairs in the batch are influencing each other). For proper cross-modal attention, you would typically want to compute the attention weights only between corresponding text and image embeddings within the same sample.

NEW QUESTION: 37

Consider a multimodal emotion recognition system that uses both facial expressions (images) and speech (audio). You want to fuse the information from these two modalities at the decision level. Which of the following techniques would be MOST suitable for decision-level fusion?

- A. Concatenate the feature vectors extracted from the images and audio, then train a single classifier.
- B. Train separate classifiers for images and audio, then average their output probabilities for each emotion class.

- C. Train separate classifiers for images and audio, then use a weighted average of their output probabilities based on the confidence scores of each classifier.
- D. Train separate classifiers for images and audio, then use the output of the image classifier as input to the audio classifier-
- E. Train a single transformer to process both images and audio in sequence.

Answer: ([SHOW ANSWER](#))

Weighted averaging allows you to give more weight to the modality that is more reliable or confident in its prediction for a given input. Simply averaging treats all modalities equally. Concatenation is feature-level fusion. The image classifier as input to audio classifier is a specific cascade approach. Using a single transformer is possible, but less common for decision fusion specifically. It is feature level fusion.

NEW QUESTION: 38

You are evaluating a multimodal model that generates descriptions for video clips. You have human ratings for the relevance, fluency, and coherence of the generated descriptions. Which statistical test is MOST appropriate for determining if there is a statistically significant difference in the median ratings for each of these criteria (relevance, fluency, coherence) between two different versions of your model?

- A. T-test
- B. ANOVA
- C. Mann-Whitney U test
- D. Kruskal-Wallis test
- E. Friedman Test

Answer: C ([LEAVE A REPLY](#))

Since you're interested in comparing the medians of the ratings and not assuming a normal distribution (which is often the case with subjective human ratings), a non-parametric test is more appropriate than a t-test or ANOVA. The Mann-Whitney U test (also known as the Wilcoxon rank-sum test) is used to compare the medians of two independent groups. The Kruskal-Wallis test is used when you have more than two groups.

NEW QUESTION: 39

Which visualization technique is suitable for representing the distribution of performance scores for different multimodal ML models over different modalities?

- A. Heatmap
- B. Histogram
- C. Box plot
- D. Pie chart

Answer: C ([LEAVE A REPLY](#))

A box plot (box-and-whisker plot) summarizes the distribution of a numeric variable - median, interquartile range, and outliers - as a single compact glyph, and critically, multiple box plots can

be placed side by side to compare distributions across categorical groupings. This makes it well suited to the scenario described:

comparing the spread and central tendency of performance scores across several models, further faceted by modality, in one readable figure. Box plots make skew, variance, and outlier prevalence immediately comparable across groups in a way a single summary statistic (like mean accuracy) cannot.

A histogram (B) shows the distribution of a single variable well but does not scale cleanly to side-by-side comparison across many model/modality combinations without becoming visually cluttered. A heatmap (A) is excellent for showing a matrix of values (e.g., mean score per model × modality pair) but represents point estimates, not distributions - it cannot convey variance or spread. A pie chart (D) is inappropriate for any continuous performance metric.

In practice, a violin plot - which overlays a kernel density estimate on the box plot's summary statistics - is often preferred when the underlying distribution's shape (e.g., bimodality) matters, but among the given options, the box plot is the correct choice for distributional comparison across groups.

Reference: Data Analysis and Visualization domain - comparative distribution visualization, box plots vs. heatmaps.

NEW QUESTION: 40

You are working with a multimodal generative model that combines text and image inputs. The model's performance is suboptimal when generating images conditioned on complex text descriptions. Which data analysis technique would be MOST effective in identifying the root cause of this issue?

- A. Calculating the average image resolution in the dataset.
- B. Performing a sentiment analysis of the text descriptions to identify potential biases.
- C. Analyzing the correlation between the complexity of the text descriptions (e.g., number of words, sentence structure) and the quality of the generated images.
- D. Measuring the frequency of different objects appearing in the images.
- E. Calculating the mean pixel intensity of the images.

Answer: (SHOW ANSWER)

Analyzing the correlation between text complexity and image quality directly investigates whether the model struggles with more intricate text descriptions. Sentiment analysis is relevant for identifying biases, but doesn't address the complexity issue directly. The average image resolution and pixel intensity are general image statistics and not specific to the multimodal problem. Object frequency in images can be useful but less direct than correlating text complexity with image quality.

NEW QUESTION: 41

You are tasked with building a system that generates realistic images based on both textual descriptions and a semantic segmentation map. The segmentation map provides spatial

information about the objects present in the scene. Which of the following generative architectures is MOST appropriate for this multimodal task?

- A. Vanilla Generative Adversarial Network (GAN)
- B. Variational Autoencoder (VAE)
- C. Conditional Generative Adversarial Network (cGAN) with both text and segmentation map as conditions.
- D. Autoregressive model like PixelCNN
- E. Diffusion model without conditioning

Answer: [\(SHOW ANSWER\)](#)

A Conditional GAN (cGAN) is the MOST suitable architecture. cGANs allow you to condition the image generation process on additional information, such as text and segmentation maps. By providing both modalities as conditions, the generator can learn to create images that are consistent with both the textual description and the spatial layout defined by the segmentation map. Vanilla GANs and VAEs don't offer explicit conditioning mechanisms. Autoregressive models can generate high-quality images, but they don't easily accommodate multimodal inputs like text and segmentation maps. A diffusion model without conditioning does not have the capacity to generate images from multimodal prompts.

NEW QUESTION: 42

When using prompt engineering with text-to-image models, which of the following techniques are most effective in improving the fidelity and relevance of generated images to the input text?

- A. Using vague and open-ended prompts to encourage creative variations.
- B. Using highly specific and detailed prompts, including attributes, style, and composition.
- C. Using negative prompts to explicitly exclude undesirable elements from the generated image.
- D. Focusing solely on the main subject of the image, omitting any contextual details.
- E. Using a combination of highly specific prompts and negative prompts.

Answer: B,C,E [\(LEAVE A REPLY\)](#)

Effective prompt engineering involves providing the model with enough specific details to understand the desired image attributes, style, and composition. Negative prompts help refine the output by explicitly excluding unwanted elements, leading to improved fidelity and relevance. Vague prompts are less effective, and omitting context can lead to undesirable or unexpected results.

NEW QUESTION: 43

During data analysis for a multimodal A1 project involving image and text data, you discover that the image dataset contains a large number of blurry or low-resolution images. The text data, however, is relatively clean and well-structured. What is the BEST approach to mitigate the impact of the noisy image data on the overall model performance?

- A. Discard the blurry and low-resolution images from the dataset to ensure data quality.
- B. Apply image enhancement techniques such as sharpening and super-resolution to improve the quality of the blurry images.

- C. Increase the weight of the text data during model training to compensate for the noisy image data.
- D. Train the model on the noisy image data without any preprocessing or data augmentation.
- E. Use a combination of image enhancement techniques and robust loss functions that are less sensitive to noisy data.

Answer: E ([LEAVE A REPLY](#))

A combination of image enhancement and robust loss functions provides the best approach. Image enhancement techniques can improve the quality of the blurry images, making them more informative for the model. Robust loss functions, such as Huber loss or Tukey's biweight loss, are less sensitive to outliers and noisy data, which can further mitigate the impact of the remaining noise. Discarding data (A) reduces the dataset size. Increasing the weight of text data (C) may lead to the model ignoring visual information. Training on raw noisy data (D) will severely impact the model's ability to learn correct mappings.

NEW QUESTION: 44

You're tasked with building a system that can generate realistic images from text descriptions and, conversely, generate accurate text descriptions from images. You decide to use a GAN (Generative Adversarial Network) architecture, but need to handle both modalities effectively. What GAN variant would be MOST suitable for this bi-directional multimodal task?

- A. Vanilla GAN
- B. Conditional GAN (cGAN)
- C. CycleGAN
- D. Deep Convolutional GAN (DCGAN)
- E. Super-Resolution GAN (SRGAN)

Answer: C ([LEAVE A REPLY](#))

CycleGAN is designed for unpaired image-to-image translation. In this scenario, it can be adapted to translate between the image and text modalities without requiring paired data. One generator learns to generate images from text, while another learns to generate text from images. Cycle consistency ensures that translating an image to text and then back to an image results in an image similar to the original. Vanilla GAN, cGAN, and DCGAN are not inherently designed for bi-directional translation between modalities without paired data. SRGAN is for image super-resolution.

NEW QUESTION: 45

You are working on a project that involves analyzing customer reviews which contains the following dataset: 1. customer_id(categorical) 2. customer_review(text) 3. product_image(image) 4. video_of_product_usage(video) What is the best way to handle and address the problem of skewness across each modalities?

- A. Treat all modalities with equal weights during model training, ignoring potential skewness issues.

- B.** Balance the dataset by oversampling under-represented data points within each modality independently.
- C.** Apply modality-specific weighting schemes that assign higher weights to modalities with less representation.
- D.** Design a loss function that explicitly penalizes the model for being biased towards dominant modalities.
- E.** Do nothing about the skewness, as the model will learn to adapt to the imbalanced data distribution.

Answer: ([SHOW ANSWER](#))

Addressing skewness is crucial for preventing the model from being biased towards dominant modalities. Oversampling, modality-specific weighting, and a biased-aware loss function are all effective strategies for mitigating this problem.

NEW QUESTION: 46

Explain the role of Tensor Cores and mixed-precision training (e.g., using FP16 or bfloat16) in accelerating the training of large generative AI models.

- A.** Tensor Cores perform specialized matrix multiplications optimized for lower-precision data types, enabling faster computation and reduced memory footprint.
- B.** Mixed-precision training allows using lower precision for forward and backward passes but keeps weights and gradients in higher precision to maintain stability.
- C.** Tensor Cores are only useful for inference, not training.
- D.** Mixed-precision training guarantees the same convergence behavior as full-precision training.
- E.** A and B.

Answer: **E** ([LEAVE A REPLY](#))

Tensor Cores are designed to accelerate matrix multiplication, the core operation in deep learning, using lower precision data types. Mixed-precision training leverages this by using lower precision for the bulk of the computation, while maintaining higher precision for critical variables to avoid instability. Tensor Cores are used both for training and inference.

Valid NCA-GENM Dumps shared by PrepPdf.com for Helping Passing NCA-GENM Exam! PrepPdf.com now offer the **newest NCA-GENM exam dumps**, the PrepPdf.com NCA-GENM exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com NCA-GENM dumps with Test Engine here:
<https://www.preppdf.com/NVIDIA/NCA-GENM-prepaway-exam-dumps.html> (**58 Q&As Dumps**, **40%OFF Special Discount: Exam-Tests**)

NEW QUESTION: 47

Consider the following Python code snippet utilizing the Hugging Face Transformers library for multimodal processing. The objective is to perform visual question answering (VQA). Assume

'image' is a PIL Image object and 'question' is a string. However, the code is incomplete. Choose the options to complete the code.

```
model_name = 'Salesforce/blip-vqa-base'
processor = AutoProcessor.from_pretrained(model_name)
model = AutoModelForQuestionAnswering.from_pretrained(model_name)
inputs = processor(image, question, return_tensors='pt')
outputs = model(inputs)
```

A. `answer = processor.decode(outputs.logits.argmax(-1))`

B.

```
model_name = 'Salesforce/blip-vqa-base'
processor = AutoProcessor.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)
inputs = processor(images=image, text=question, return_tensors='pt')
outputs = model.generate(inputs)
answer = processor.decode(outputs[0], skip_special_tokens=True)
```

```
model_name = 'Salesforce/blip-vqa-base'
processor = AutoProcessor.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
inputs = processor(images=image, text=question, return_tensors='pt')
outputs = model.generate(inputs)
answer = processor.decode(outputs[0], skip_special_tokens=True)
```

C.

```
model_name = 'Salesforce/blip-vqa-base'
processor = AutoProcessor.from_pretrained(model_name)
model = AutoModel.from_pretrained(model_name)
inputs = processor(images=image, text=question, return_tensors='pt')
outputs = model(inputs)
```

D. `answer = outputs.logits.argmax(-1)`

```
model_name = 'Salesforce/blip-vqa-base'
processor = AutoProcessor.from_pretrained(model_name)
model = AutoModelForImageClassification.from_pretrained(model_name)
inputs = processor(images=image, text=question, return_tensors='pt')
outputs = model(inputs)
answer = processor.decode(outputs[0], skip_special_tokens=True)
```

E.

Answer: C (LEAVE A REPLY)

The correct code uses 'AutoModelForSeq2SeqLM' because BLIP (used in the example) is a sequence-to-sequence model. The processor correctly handles the image and text, and 'model.generate' produces the answer which is then decoded. 'AutoModelForQuestionAnswering' is not a generic class and won't work correctly with BLIP without additional adaptation.

NEW QUESTION: 48

In ML applications, which machine learning algorithm is commonly used for creating new data based on existing data?

- A. Decision tree
- B. Support vector machine (SVM)
- C. K-means clustering
- D. Generative adversarial network (GAN)

Answer: D ([LEAVE A REPLY](#))

GANs are purpose-built generative models: as covered in the previous question, the generator component learns the underlying distribution of a training dataset and produces new synthetic samples that resemble it - new images, audio, or other data types that did not exist in the original dataset but are statistically consistent with it. This generative capability is GAN's defining characteristic and the reason it is the correct answer among the options given, distinguishing it from the other three algorithms, all of which are fundamentally discriminative or unsupervised techniques rather than generative ones.

Decision trees (A) and support vector machines (B) are supervised discriminative algorithms - they learn a decision boundary or a set of rules to classify or predict outputs from inputs, with no mechanism for producing novel data samples resembling a training distribution. K-means clustering (C) is unsupervised but serves a partitioning function, grouping existing data points into clusters based on similarity - it identifies structure in data that already exists rather than synthesizing new data points that didn't exist before.

It's worth noting GANs are one of several generative model families (alongside variational autoencoders and diffusion models, both covered elsewhere in this set) - among the four options presented here, however, GAN is the only one designed for generation at all, making this a comparatively direct elimination once the discriminative-vs-generative distinction is applied.

Reference: Core Machine Learning and AI Knowledge domain - generative vs. discriminative algorithms, GANs.

NEW QUESTION: 49

Consider a scenario where you are developing a multimodal A1 system to translate sign language videos into text. The system utilizes a CNN for processing video frames and an RNN for generating the text sequence. During evaluation, you observe that the system struggles to accurately translate signs that involve complex hand movements or subtle facial expressions. What are the MOST effective strategies to improve performance in this specific scenario? (Select TWO)

- A. Reduce the frame rate of the input video to decrease computational load.
- B. Increase the depth of the CNN to capture more fine-grained visual features.
- C. Incorporate a 3D CNN architecture to better capture spatial-temporal information from the video.
- D. Replace the RNN with a simpler feed forward neural network for faster processing.

E. Augment the training data with variations of the sign language videos, including different camera angles, lighting conditions, and signers.

Answer: C,E ([LEAVE A REPLY](#))

3D CNNs are designed to capture spatial-temporal information, which is essential for recognizing complex hand movements. Data augmentation helps the model generalize to different conditions and variations in signing style. Reducing frame rate (A) would likely worsen performance.

Replacing the RNN with a feed forward network (D) would remove the ability to model sequential information, which is critical for translation. Increasing the depth of the CNN (B) might help, but a 3D CNN is more directly suited to the task.

NEW QUESTION: 50

You are tasked with building a system that can generate captions for images. You want to use a transformer-based model. During inference, you notice that the model tends to generate repetitive captions. Which of the following decoding strategies could you use to mitigate this issue?

- A. Greedy decoding
- B. Beam search with a high beam width
- C. Beam search with a length penalty
- D. Random sampling
- E. Top-k sampling

Answer: C,E ([LEAVE A REPLY](#))

Greedy decoding often leads to repetitive sequences because it always chooses the most likely next word. Beam search, especially with a high beam width, can also lead to repetition if the highest-probability sequences contain repeated phrases. Beam search with a length penalty encourages the model to generate longer and more diverse captions. Random sampling and Top-k sampling introduce randomness into the decoding process, which can help to break out of repetitive loops and create more varied captions.

NEW QUESTION: 51

Given the following Python code snippet using Pandas, which is intended to filter rows where the 'price' column is greater than 100 and the 'quantity' column is less than 5, identify the correct approach to achieve this:

- A. `df[(df.price > 100) | (df.quantity < 5)]`
- B. `df.query('price > 100 and quantity < 5')`
- C. `df.filter(items=['price > 100', 'quantity < 5'])`
- D. `df[(df['price'] > 100) & (df['quantity'] < 5)]`
- E. `df[(df.price > 100) and (df.quantity < 5)]`

Answer: ([SHOW ANSWER](#)**)**

Option B uses the 'query' method for concise filtering. Option D correctly uses boolean indexing with the (and) operator within square brackets. Option E is syntactically incorrect because it uses the 'and' keyword, which is meant for single boolean values, not Pandas Series. Option A does

logical OR which is not what we intend to filter. Option C is incorrect since filter method expects list of column names.

NEW QUESTION: 52

You have trained a multimodal model to generate descriptions for recipes, using images of the finished dish as one modality and a list of ingredients as another. When evaluating the generated descriptions, you notice the descriptions are factually correct in terms of ingredients but often fail to capture the stylistic nuances or tone of professionally written recipes. Which evaluation strategy would provide the most insightful feedback on this aspect of the model's performance?

- A. Calculate the BLEU score between the generated descriptions and a set of reference recipes.
- B. Compute the ROUGE score between the generated descriptions and a set of reference recipes.
- C. Use a pre-trained language model to assess the fluency and coherence of the generated descriptions.
- D. Conduct a human evaluation where individuals rate the generated descriptions based on factors such as informativeness, clarity, and stylistic appropriateness.
- E. Measure the perplexity of the generated descriptions using a language model trained on a corpus of recipe books.

Answer: D ([LEAVE A REPLY](#))

Human evaluation is the most reliable way to assess subjective aspects of the generated text, such as stylistic appropriateness. While automated metrics like BLEU, ROUGE, and perplexity can provide insights into factual accuracy and fluency, they do not capture the nuances of style and tone as effectively as human judgment. Therefore, the correct answer is (D).

NEW QUESTION: 53

You are tasked with building a multimodal generative AI model to create marketing content from product images and descriptions. The image encoder uses a pre-trained ResNet50 model, and the text encoder uses a pre-trained BERT model. After initial training, the generated content frequently misinterprets the image. Which of the following strategies is MOST effective in improving the model's ability to correctly interpret the image within the multimodal context?

- A. Freeze the weights of both the ResNet50 and BERT models to prevent overfitting.
- B. Fine-tune the ResNet50 model with a dataset of images specifically related to the product domain, using a contrastive loss function that encourages representations of images and corresponding text to be close in the embedding space.
- C. Increase the learning rate for the BERT model to prioritize text-based information.
- D. Replace ResNet50 with a simpler image encoder like a shallow CNN to reduce computational complexity.
- E. Decrease the batch size during training.

Answer: B ([LEAVE A REPLY](#))

Fine-tuning ResNet50 with a relevant image dataset and a contrastive loss function directly addresses the issue of misinterpreting the image. Freezing weights prevents learning, increasing

BERT's learning rate imbalances the model, and a simpler image encoder might lose crucial image details. Decreasing batch size can improve generalization but isn't the primary solution for image misinterpretation.

NEW QUESTION: 54

Consider the following Python code snippet using Triton Inference Server's Python client. The code intends to send a request to a model that expects two input tensors: 'input_image' (shape: [1, 3, 224, 224], datatype: FP32) and 'input_text' (shape: [1,], datatype: BYTES). Identify potential issues in this code that could prevent successful inference.

- A. The function is used incorrectly; it should directly accept the Triton datatype string (e.g., 'FP32').
- B. The input data for 'input_text' needs to be encoded to bytes using UTF-8 encoding before being passed to Triton.
- C. The data is not converted to the appropriate NumPy datatype before being sent to Triton.
- D. The model name and input/output names must be specified, but they are missing in the code.
- E. All of the above.

Answer: E ([LEAVE A REPLY](#))

All the mentioned issues (A, B, C, D) can prevent successful inference. requires correction. The input text requires explicit byte encoding. Converting input_image to the correct numpy data type. Specifying the model name and input/output names is important for triton to understand the request. If all of these requirements are not met, the triton request will fail.

NEW QUESTION: 55

What is the significance of A/B testing in ML software engineering?

- A. A/B testing is used to measure the impact of changes in the user interface of a ML application.
- B. A/B testing helps in optimizing the hyperparameters of a machine learning model.
- C. A/B testing is irrelevant in ML software engineering.
- D. A/B testing helps in evaluating the performance and effectiveness of different machine learning models.

Answer: D ([LEAVE A REPLY](#))

A/B testing in an ML engineering context is a controlled experimental methodology where two (or more) model variants - for example, a current production model (A) and a candidate replacement (B) - are deployed simultaneously to randomly assigned, statistically comparable segments of live traffic, and their real- world performance is compared on business or task-relevant metrics (conversion rate, click-through rate, task accuracy, latency-adjusted engagement). This provides causal evidence of which model performs better under actual production conditions, which offline evaluation on static held-out datasets cannot fully capture, since production data distributions shift and downstream user behavior interacts with model outputs in ways offline metrics miss.

Option A narrows A/B testing incorrectly to UI changes alone; while A/B testing originated in and remains common for UI/UX experimentation, its application in ML engineering explicitly extends to comparing model versions, algorithms, and feature sets - not just interface elements. Option B

misattributes a distinct methodology (systematic hyperparameter search, covered in the previous question) to A/B testing, which evaluates already-trained model variants rather than searching a hyperparameter space. Option C is simply incorrect - A/B testing is a cornerstone of responsible ML deployment and MLOps practice, gating rollout decisions before full production release. Reference: Experimentation domain - A/B testing, online evaluation, controlled experimentation in production ML.

NEW QUESTION: 56

You're tasked with building a model that can generate recipes from images of food. You decide to use a Variational Autoencoder (VAE) architecture. What would be a suitable loss function combination for this task, considering both reconstruction accuracy and recipe relevance?

- A. KL divergence loss only-
- B. Reconstruction loss (MSE) between the input image and the decoded image only
- C. Reconstruction loss (MSE) + KL divergence loss + Cross-entropy loss between the generated recipe and a plausible recipe given the generated image embedding-
- D. Reconstruction loss (MSE) + Perceptual loss (based on a pre-trained image classifier) only
- E. KL divergence loss + Cosine Similarity loss between the generated image embedding and a text embedding of a random recipe.

Answer: C ([LEAVE A REPLY](#))

The Reconstruction loss ensures the generated image is similar to the input. KL divergence enforces a smooth latent space. The Cross-entropy loss ensures the generated recipe is relevant to the decoded image. Perceptual loss, while helpful for image quality, doesn't directly address recipe relevance. Using a text embedding of a random recipe would not guide the model towards generating relevant recipes.

NEW QUESTION: 57

You are working on a project that involves generating realistic images of furniture based on textual descriptions. The input data consists of text descriptions and a small dataset of existing furniture images. Which data augmentation techniques would be MOST effective in improving the quality and diversity of the generated images?

- A. Randomly cropping and rotating the existing furniture images.
- B. Using generative adversarial networks (GANs) to generate new furniture images from the existing dataset.
- C. Synthesizing new text descriptions using paraphrasing and back-translation techniques.
- D. Combining A, B, and C.
- E. Focusing solely on increasing the size of the text dataset and ignoring image augmentation.

Answer: D ([LEAVE A REPLY](#))

Combining all techniques provides the best results. Image augmentations like cropping and rotation increase the variance of the image data. GANs create entirely new images, and text augmentation enhances the diversity of the input descriptions. Focusing only on one modality will likely limit the model's performance.

NEW QUESTION: 58

You're tasked with building a system that generates personalized exercise recommendations based on user's text descriptions of their fitness goals and images of their current physical condition. Due to privacy concerns, you cannot directly access the user's raw images or text after initial processing. What technique can allow you to continue to train the model while respecting these privacy constraints?.

- A. Federated Learning
- B. Transfer Learning
- C. Data Augmentation
- D. Generative Adversarial Networks (GANs)
- E. Reinforcement Learning

Answer: A ([LEAVE A REPLY](#))

Federated learning allows training a model across multiple decentralized devices or servers holding local data samples, without exchanging them. This is perfect for privacy-sensitive scenarios as the raw data remains on the user's device. Transfer learning relies on pre-trained models, data augmentation modifies existing data, GANs generate new data (but still require initial data access), and reinforcement learning optimizes actions through interaction with an environment.

NEW QUESTION: 59

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 ([LEAVE A REPLY](#))

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.

NEW QUESTION: 60

Consider the following code snippet used within a U-Net architecture. What is its purpose?

```
torch.cat([up, skip], dim=1)
```

- A. It performs an element-wise addition of the 'up' and 'skip' tensors.
- B. It concatenates the 'up' and 'skip' tensors along the channel dimension.
- C. It multiplies the 'up' and 'skip' tensors element-wise.
- D. It performs a matrix multiplication between the 'up' and 'skip' tensors.
- E. It subtracts the 'skip' tensor from the 'up' tensor.

Answer: B ([LEAVE A REPLY](#))

The 'torch.cat([up, skip], dim=1)' function concatenates two tensors, 'up' and 'skip', along the channel dimension (dim=1). In the context of a U-Net, 'up' represents the upsampled feature map from the decoder path, and 'skip' represents the corresponding feature map from the encoder path. Concatenating them allows the decoder to combine both coarse-grained and fine-grained information for better image reconstruction.

NEW QUESTION: 61

You are building a multimodal generative AI model that combines text, images, and audio. You notice that the model performs well on text and images but struggles with audio, particularly in noisy environments. Which of the following strategies would be MOST effective in improving the model's performance with audio data?

- A. Increase the learning rate for the audio modality during training.
- B. Reduce the dimensionality of the audio features to simplify the learning task.
- C. Apply data augmentation techniques specifically designed for audio, such as adding noise or varying the speed and pitch.
- D. Decrease the weight of the audio modality in the loss function.
- E. Use transfer learning by pre-training the audio component of the model on a large audio dataset.

Answer: C,E ([LEAVE A REPLY](#))

Data augmentation (C) increases the robustness of the model to variations in audio, including noise. Transfer learning (E) allows the model to leverage knowledge from a large, pre-existing audio dataset, improving its initial performance.

Valid NCA-GENM Dumps shared by PrepPdf.com for Helping Passing NCA-GENM Exam! PrepPdf.com now offer the **newest NCA-GENM exam dumps**, the PrepPdf.com NCA-GENM exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com NCA-GENM dumps with Test Engine here:

NEW QUESTION: 62

You are training a multimodal generative A1 model for image captioning. After initial training, you observe that the model excels at describing common objects but struggles with nuanced details and rare objects. Which of the following performance optimization strategies would be MOST effective in addressing this issue?

- A. Increase the batch size during training to improve GPU utilization.
- B. Implement a custom loss function that penalizes inaccuracies in describing rare objects more heavily.
- C. Apply early stopping to prevent overfitting to the common objects.
- D. Reduce the learning rate to fine-tune the model on the existing dataset.
- E. Increase the number of layers in the encoder network.

Answer: ([SHOW ANSWER](#))

Implementing a custom loss function is the most effective strategy because it directly addresses the model's weakness by focusing on accurate descriptions of rare objects. Increasing batch size improves training speed but not necessarily accuracy. Early stopping prevents overfitting, but doesn't specifically target the issue of rare object recognition. Reducing the learning rate might help with fine-tuning, but not as effectively as a targeted loss function. Increasing the number of layers may increase complexity but not guarantee better performance on rare objects.

NEW QUESTION: 63

You're training a multimodal model to generate 3D models from text descriptions. The models are evaluated using Intersection over Union (IOU) between the generated and ground truth 3D models. During evaluation, you observe perfect IOU scores on some samples, but visual inspection reveals significant discrepancies. What is the MOST likely cause for this, and what can be done to correct the process?

- A. The model is overfitting, resulting in near-perfect reconstruction of a subset of training samples. Reduce the model's capacity.
- B. There is a data leakage issue, where some of the test data is present in the training data. Ensure that training and test data are completely disjoint.
- C. The IOU calculation is being performed incorrectly, or there is a bug in the evaluation code. Verify the IOU implementation.
- D. The text descriptions are too simple. Use more complex text prompts to prevent overfitting.
- E. IOU is an inherently flawed metric for evaluating 3D models and needs to be replaced by Chamfer distance.

Answer: C ([LEAVE A REPLY](#))

Perfect IOU scores with visual discrepancies strongly suggest a problem with the IOU calculation itself (C). Data leakage (B) or overfitting (A) are possibilities, but a bug in the IOU implementation is more likely given the perfect scores. Text complexity (D) doesn't explain perfect scores with

visual errors. IOU is a valid metric, and it could be supplemented with chamfer distance, but if IOU gives perfect scores with visual discrepancies, then the immediate action needed is to verify IOU implementation. Thus, the best option is C.

Valid NCA-GENM Dumps shared by PrepPdf.com for Helping Passing NCA-GENM Exam!

PrepPdf.com now offer the **newest NCA-GENM exam dumps**, the PrepPdf.com NCA-GENM exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com NCA-GENM dumps with Test Engine here:

<https://www.preppdf.com/NVIDIA/NCA-GENM-prepaway-exam-dumps.html> (**58** Q&As Dumps,

40%OFF Special Discount: **Exam-Tests**)