Contents

WEEK 9: FROM SUPERVISED TO GENERATIVE LEARNING

1. Beyond Supervised Learning

1.1 Limitations of Labeled Data

Limitations of Supervised Learning Paradigms: A Technical Analysis

The epistemological foundations of supervised learning methodologies are intrinsically constrained by their dependence on annotated datasets, wherein input-output pairs $(x_i, y_i) \in X \times Y$ are manually curated to form the training distribution $P(X, Y)$. While this framework has demonstrated empirical success across numerous domains, particularly in discriminative tasks where $P(Y|X)$ is of primary interest, the approach encounters fundamental scaling limitations that warrant rigorous examination.

The annotation process presents both economic and methodological challenges in the construction of large-scale supervised datasets. The required human expertise introduces significant temporal and financial costs that scale linearly (or in some cases super-linearly) with dataset size, creating an upper bound on practical dataset construction. This limitation becomes particularly acute in specialized domains such as medical image analysis, where expert annotators (e.g., board-certified radiologists) must perform pixel-wise segmentation of anatomical structures, or in computational linguistics, where syntactic parse trees and semantic role labels require expert linguistic knowledge for accurate annotation.

From an information-theoretical perspective, the supervised paradigm inherently suffers from the curse of dimensionality when attempting to model complex manifolds in high-dimensional feature spaces. The number of required labeled examples often grows exponentially with the intrinsic dimension of the target function, leading to a fundamental tension between model capacity and dataset scalability. This theoretical limitation manifests practically in domains such as medical imaging, where the feature space dimensionality can exceed 10^7 dimensions for high-resolution 3D scans, necessitating novel approaches that can leverage unlabeled data more effectively through self-supervised or semi-supervised learning frameworks.

Scale limitations become particularly apparent when dealing with modern deep learning systems, which often require millions of labeled examples to achieve high performance. The manual labeling process simply cannot keep pace with the growing demand for training data. This becomes even more challenging when dealing with rare events or edge cases, where obtaining labeled examples is inherently difficult due to their infrequent occurrence.

1.2 The Generative Alternative

The emergence of generative approaches offers a promising alternative to traditional supervised learning. Instead of learning direct mappings from inputs to outputs, generative models learn to understand and reproduce the underlying data distribution. This fundamental shift in approach allows models to learn from unlabeled data, which is typically abundant and easier to obtain. By modeling the data distribution itself, these models can capture complex patterns and relationships without requiring explicit labels.

Self-supervised learning has emerged as a powerful paradigm within the generative framework. In this approach, the model creates its own supervisory signals from the raw data, effectively learning from the data's inherent structure. For example, in text generation, a model might learn by predicting missing words in sentences, using the natural context as supervision. This allows the model to learn rich representations without requiring manual annotations.

The distinction between implicit and explicit modeling represents another crucial aspect of generative approaches. Explicit models, such as autoregressive models, directly learn the probability distribution of the data. In contrast, implicit models, such as GANs or diffusion models, learn to generate data without explicitly modeling the full probability distribution. This flexibility in modeling approaches has led to breakthroughs in various domains, from image synthesis to text generation.

2. Diffusion Models: Core Concepts

2.1 Forward Process

The forward process in diffusion models represents a systematic approach to gradually destroying structure in data. This process can be understood as adding noise to data in small, controlled steps until the data becomes pure noise. Imagine a photograph slowly fading into static - each step slightly blurs the image until the original is completely obscured. This process is carefully designed to be reversible, making it possible to learn how to reconstruct the original data.

The noise scheduling in diffusion models is a critical component that determines how quickly and in what manner information is destroyed. This schedule typically follows a variance curve that starts slow and accelerates, allowing the model to learn the destruction process at different scales. The careful balance of this schedule ensures that information is lost gradually enough for the reverse process to learn effectively, while still reaching a state of pure noise within a reasonable number of steps.

The mathematical foundation of the forward process relies on Markov chains, where each step depends only on the immediate previous state. This property is crucial as it simplifies the learning process and makes the model more tractable. The Markov property ensures that at each step, the model only needs to understand how to add a small amount of noise, rather than keeping track of the entire history of changes.

2.2 Reverse Process

The reverse process is where the magic of diffusion models truly happens. Starting from pure noise, the model learns to gradually reconstruct the original data distribution by reversing the forward process step by step. This denoising process requires the model to understand the underlying structure of the data at different levels of abstraction. At each step, the model must predict what the slightly less noisy version of the input should look like.

Score matching plays a central role in the reverse process. The model learns to estimate the gradient of the log probability density (the score) of the data distribution at each noise level. This approach allows the model to learn how to move noisy data points toward regions of higher probability in the true data distribution. The beauty of this approach is that it doesn't require explicit modeling of the entire probability distribution, making it more computationally tractable.

Time conditioning is a crucial aspect of the reverse process. The model must understand what noise level it's working with to properly denoise the data. This is typically achieved by conditioning the model on a time step or noise level parameter. This conditioning allows the same model to handle different stages of the

denoising process, from very noisy inputs to almost-clean data, effectively learning a continuous spectrum of denoising behaviors.

2.3 Applications

Diffusion models have shown remarkable success in image generation tasks, producing some of the highest quality synthetic images to date. The gradual nature of the denoising process allows these models to capture fine details while maintaining global coherence. This has led to applications in various domains, from artistic creation tools to medical image synthesis.

In the realm of text-to-text tasks, diffusion models are beginning to show promise, although their application is less straightforward than in the image domain. The discrete nature of text presents unique challenges, but researchers have developed clever adaptations of the diffusion process for sequential data. These models can be used for tasks like text style transfer, paraphrasing, and even machine translation.

Cross-modal generation represents one of the most exciting applications of diffusion models. By learning the relationship between different modalities (such as text and images), these models can perform tasks like text-to-image generation or image-to-text description. The gradual nature of the diffusion process allows for fine-grained control over the generation process, making it possible to create highly detailed and accurate cross-modal translations.

3. Self-Supervised Learning

3.1 Masked Prediction Tasks

Masked prediction tasks represent one of the most successful approaches to self-supervised learning, pioneered by models like BERT in natural language processing. The core idea is elegantly simple yet powerful: deliberately hide parts of the input data and train the model to reconstruct them. This approach leverages the natural structure and redundancy present in data to create a supervised learning signal without requiring external labels. For example, in text, a model might need to predict a masked word using only the surrounding context, forcing it to develop a deep understanding of language patterns and semantics.

The concept of masking extends beyond text to other domains such as computer vision. In image patch prediction, portions of images are masked out, and the model must reconstruct the missing regions. This task requires understanding complex visual patterns, spatial relationships, and the natural statistics of images. The success of models like ViT (Vision Transformer) demonstrates how masking strategies can be effectively adapted across different domains while maintaining their fundamental principles.

The process of corruption and reconstruction serves as a powerful learning mechanism. By varying the amount and pattern of masking, models can learn to understand data at multiple scales and levels of abstraction. This approach has proven particularly effective because it forces the model to develop robust internal representations that capture the underlying structure of the data, rather than merely memorizing surface-level patterns.

3.2 Contrastive Learning

Contrastive learning introduces a fundamentally different approach to self-supervised learning by focusing on the relationships between different views or instances of data. The core principle involves creating positive pairs (different views of the same instance) and negative pairs (views from different instances), then training the model to recognize these relationships. For example, two different augmentations of the same image should be recognized as similar, while augmentations of different images should be distinguished from each other.

The concept of positive and negative pairs provides a powerful framework for learning meaningful representations. By carefully designing what constitutes a "positive" pair, we can encode our prior knowledge about what features or properties should be considered invariant or important. This approach has led to state-of-the-art results in various domains, particularly in computer vision where techniques like SimCLR have demonstrated the effectiveness of contrastive learning.

Momentum contrast represents an important advancement in contrastive learning techniques. This approach maintains a dynamic dictionary of encoded keys, which allows for comparison against a much larger set of negative examples without requiring an excessive batch size. The momentum update mechanism ensures stable learning while maintaining computational efficiency, making it possible to scale contrastive learning to larger datasets and more complex tasks.

3.3 Connection to Diffusion

The relationship between self-supervised learning and diffusion models reveals interesting theoretical and practical connections. Denoising as a form of self-supervision represents a natural bridge between these approaches. In diffusion models, the denoising process can be viewed as a continuous form of masked prediction, where the "mask" is Gaussian noise added at various scales. This perspective helps unify our understanding of different self-supervised learning approaches.

The concept of learning without labels takes on new meaning when viewed through the lens of diffusion models. Rather than requiring explicit supervision, these models learn from the intrinsic structure of the data itself, similar to how contrastive learning leverages data relationships. This connection highlights how different self-supervised approaches can complement each other and contribute to our understanding of unsupervised learning.

The quality of representations learned through these approaches often exceeds what can be achieved through traditional supervised learning. This is particularly evident when examining the internal representations of diffusion models, which must capture complex hierarchical structures to successfully denoise data. The gradual nature of the diffusion process allows the model to learn representations at multiple scales, from fine details to global structure, resulting in rich and useful feature hierarchies.