Topic: Text-to-image diffusion models in generative ai

Survey & papers:

Text-to-image diffusion models in generative ai

Diffusion Models Beat GANs on Image Synthesis

High-Resolution Image Synthesis with Latent Diffusion Models (LDM)

Quick look on this topic:

۱ .سیر تکامل الگوریتم‌ها: از فضای پیکسل تا فضای پنهان و هدایت‌گرها مسیر توسعه مدل‌های تولید تصویر ابتدا تحت سلطه شبکه‌های GAN و مدل‌های اتورگرسیو بود، اما مدل‌های انتشار (Diffusion Models) با ارائه کیفیت بهتر و پوشش توزیع دیتای قوی‌تر، جایگزین آن‌ها شدند .مقاله *Dhariwal & Nichol*نشان داد که با بهبود معماری U-Net و معرفی روش **Classifier Guidance**(هدایت‌گر طبقه‌بند)، مدل‌های انتشار می‌توانند در کیفیت تصاویر از GAN ها پیشی بگیرند .در این روش، از گرادیان‌های یک کلاسیفایر برای هدایت فرآیند حذف نویز (Denoising) به سمت کلاس مورد نظر استفاده می‌شود که تعادلی بین تنوع و کیفیت ایجاد می‌کند .با این حال، مدل‌های پیکسلی به محاسبات سنگینی نیاز داشتند .برای حل این مشکل، مقاله *Latent Diffusion Models (LDM)* رویکرد **Latent Space**(فضای پنهان) را معرفی کرد .در این روش، به جای کار مستقیم روی پیکسل‌ها، فرآیند انتشار در یک فضای فشرده‌شده و کم‌بعد) که توسط یک Autoencoder ایجاد شده (انجام می‌شود که هزینه‌های محاسباتی را به شدت کاهش داده و امکان تولید تصاویر با وضوح بالا را فراهم می‌کند.

۲ .دیتاست‌ها و مکانیزم‌های شرطی‌سازی متنی (Conditioning) برای آموزش این مدل‌ها به دیتاست‌های عظیمی از زوج‌های "تصویر-متن" نیاز است .دیتاست‌های کلیدی مورد استفاده شامل) **ImageNet**بیشتر برای تولید بر اساس کلاس) **MS-COCO** , (برای ارزیابی همسویی متن و تصویر (و دیتاست‌های عظیم‌تری مانند **LAION-400M**هستند که برای آموزش مدل‌های قدرتمندی مثل LDM به کار رفته‌اند .برای اینکه مدل بتواند متن را بفهمد و تصویر مرتبط تولید کند، از مکانیزم‌های **Cross-Attention**استفاده می‌شود .در مدلLDM ، توصیفات متنی توسط ترنسفورمرها) مانند BERTیا (CLIP کدگذاری شده و سپس از طریق لایه‌های توجه (Attention Layers) به مدل U-Net تزریق می‌شوند تا فرآیند تولید تصویر را بر اساس متن کاربر کنترل کنند .مدل‌های پیشرفته‌تر مانند) Stable Diffusion که نسخه بزرگ‌شده LDM است (و DALL-E 2 از فضای نهان چندوجهی مدل‌هایی مثل CLIP برای همسوسازی دقیق‌تر متن و تصویر بهره می‌برند.