* VQGAN Paper zu VQGAN
* Ziel
  + Sie wollen image synthesis für high resolution Bilder machen
  + Dies mit Zusammenspiel von CNN und Transformers
  + Sie wollen irgendwie longrange dependency lernen d.h. Zusammenhänge von Elementen welche sich in einem Bild weit entfernt befinden
* Background
  + CNN haben bias gegenüber local relationships
  + Hypothesis: Locale Strukturen können mit CNN erkennt und dann mit Transformers kombiniert werden
  + Für low resolution images waren transformers besser, aber für high resolution war es infeasible
* Lernen eines codebook von Bildkomponenten
  + Sie haben encoder und decoder, welche lernen Bilder über diskretes codebook abzudecken
  + Mit Quantization zwischen encoder und decoder, welche jeden spatial code zu closest codebook entry mappt (wie bei VQVAE)
  + Damit codebook möglichst perceptually rich ist, führen sie VQ-GAN ein mit discriminator und perceptual loss
    - Perceptual loss ist der loss vom anderen Paper
  + Patch-based Discriminator versucht zwischen realen und rekonstruierten Bilder zu unterscheiden
  + Hyperparameter
    - Length Codebook Z
    - Reduction factor f
      * Wie viel kleiner sind Bilder im latent space gegenüber dem Input
        + H/f x W/f
  + Repräsentation
    - Finale Repräsentation ist eine 2D sequenz von codes, welche generiert wird indem jeder code durch dessen code index ersetzt wird
  + Ein Bild, das Text, Diagramm, Screenshot, Plan enthält.

    Automatisch generierte Beschreibung
* Image synthesis
  + Image synthesis Task ist dann eine next-index prediction von codes, welche dann die code sequenz generiert
  + Diese code sequenz wird dann mit Decoder zu einem Bild gemacht
* Wichtigster Part: Repräsentation von Bilder in latent space mit VQGAN, welches dann auch in latent diffusion verwendet wird
* Tipps
  + Beim Trainieren von VQGAN den Discriminator nicht von Anfang an hinzufügen, sondern zuerst nur auf perceptual loss und dann am Ende den Discriminator hinzufügen.

Zusammenfassung generative models

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

Fragen

* Was sind downsampling blocks
* Was ist genau reduction factor f?