Text based generation of other modalities

Alyssa Schmidt - KI in Medienanwendungen

Gliederung

- Motivation
- Grundidee
- Text-Embeddings
- CLIP: Grundlage moderner Modelle
  - Wie CLIP lernt: Contrastive Learning
  - Warum CLIP so wichtig ist
- DINO: Self-Supervised Vision Transformer
  - Wie funktioniert DINO?
- Steuerung durch Textprompts
  - Was Prompts steuern können
  - Prompt-Beispiele
- Generative Modelle für Text→Modalitäten
- Zusammenfassung

Motivation

Generative KI kann heute aus einem Textprompt:

- Bilder erzeugen
- Videos zu erzeugen
- Audio & Musik erzeugen

Kernfrage:

Wie versteht das Modell den Text – und wie nutzt es ihn, um andere Modalitäten zu erzeugen?

Grundidee

Ablauf der Textsteuerung

1. Textprompt wird eingegeben
2. Text wird tokenisiert – also in einzelne Wörter oder Teile zerlegt
3. Tokens werden zu Embeddings
4. Transformer erzeugt kontextuelle Bedeutung
5. Ergebnis: ein Textvektor (Steuervektor)
6. Generatives Modell sucht eine Ausgabe, die zu diesem Vektor passt

Wichtig:
Dieser Steuervektor ist die zentrale Kontrolleinheit bei Text-zu-Bild/Video/Audio.

Text-Embeddings

- Jeder Prompt wird in einen hochdimensionalen **Vektor** übersetzt
- Der Vektor enthält semantische Infos über:
  - Objekte
  - Eigenschaften
  - Beziehungen
  - Stil
- Dieser Vektor ist **der Steuerhebel** für die generative KI

CLIP: Grundlage moderner Modelle

- CLIP (OpenAI, 2021) = Contrastive Language-Image Pretraining

Was ist CLIP?

- CLIP ist ein Modell, das Text- und Bildinformationen in einem gemeinsamen semantischen Raum abbildet.
- CLIP ist auf **400 Millionen** Bild-Text-Paaren trainiert
- Daten stammen aus **ungefilterten Internetquellen**
- deutlich größer und vielfältiger als klassische Datensätze wie ImageNet

Wie CLIP lernt: Contrastive Learning

Grundidee:

CLIP lernt, welche Texte zu welchen Bildern gehören, indem es

Millionen ungefilterte Bild-Text-Paare aus dem Internet verwendet.

Trainingsprinzip: **Contrastive Learning**

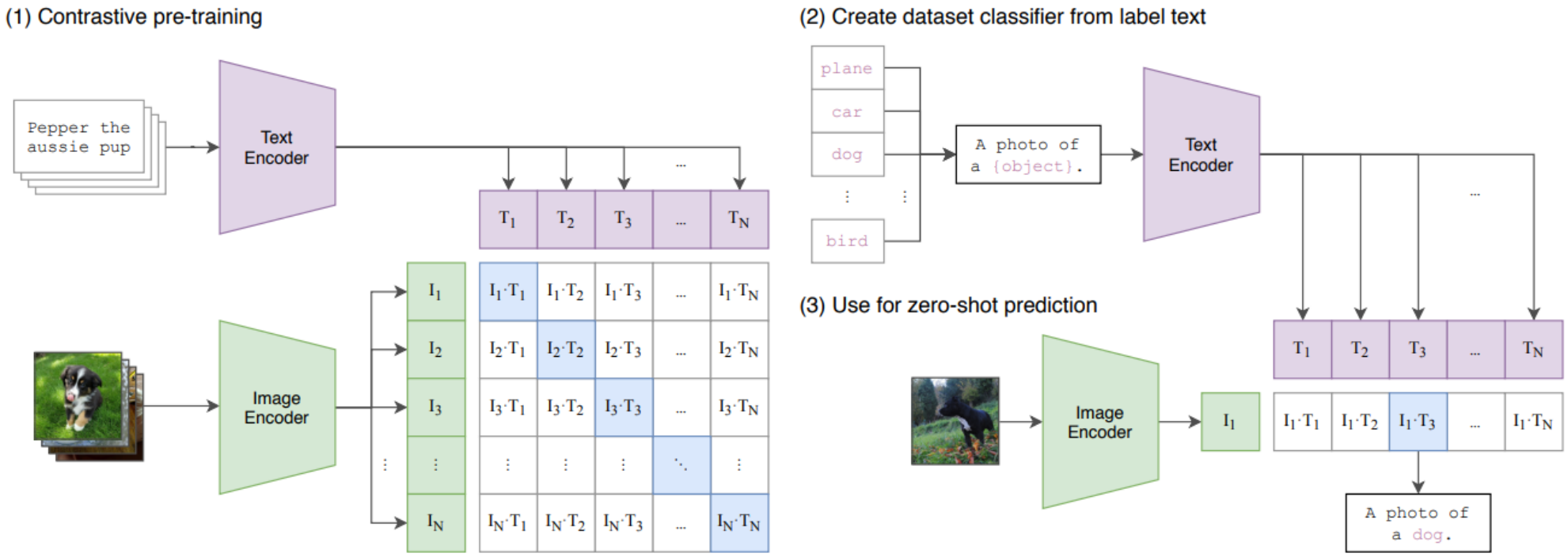
Für jeden Batch (Beispiel: 5 Bilder + 5 Texte):

- **Bildencoder** wandelt jedes Bild in einen Vektor um
- **Textencoder** wandelt jeden Text in einen Vektor um
- CLIP berechnet eine Ähnlichkeitsmatrix zwischen allen Bild-Text-Kombinationen (z. B. 5×5 Kombinationen)

Ziel:

- Die *richtigen* Paare (Bild ↔ passender Text) sollen eine **hohe Ähnlichkeit** haben
- Alle *anderen* Kombinationen sollen **niedrige Ähnlichkeit** haben

Wie CLIP lernt: Contrastive Learning



Warum CLIP so wichtig ist

CLIP ermöglicht:

- die Abbildung von Text- und Bildinhalten in einen gemeinsamen semantischen Raum
- die Bewertung der Übereinstimmung zwischen generierten Inhalten und Textprompt
- Orientierung für das Modell im latenten Raum

CLIP ist die Grundlage moderner Text-zu-Bild/Video/Audio-Modelle. Es erzeugt selbst nichts, sondern es **bewertet**, ob etwas zum Prompt passt.

DINO: Self-Supervised Vision Transformer

DINO & DINOv2 (Meta AI)

Self-Supervised Vision Transformer (ViT)

Kerngedanken:

- Lernt visuelle Repräsentationen ohne Labels
- Erkennt Objekte und Strukturen sehr zuverlässig
- Funktioniert stabil für viele Bildtypen
- Liefert starke Features für Multimodalität

Bedeutung für Text-zu-X:

- CLIP verbindet Text + Bild
- **DINO verbessert die Bildseite** (Encoder)
- In vielen modernen Multimodal-Modellen wird DINOv2 als Encoder genutzt

Wie funktioniert DINO?

DINO besteht aus zwei Netzwerken:

- Teacher-Netzwerk
- Student-Netzwerk

Beide sehen **unterschiedliche Ausschnitte des gleichen Bildes.**

Trainingsprinzip:

- Der **Teacher** erzeugt eine stabile Repräsentation seiner Bildansicht.
- Der **Student** sieht eine andere Ansicht desselben Bildes.
- Der Student versucht, die Darstellungsweise des Teachers **nachzuahmen** (zu matches).
- Das Teacher-Netzwerk wird durch ein **EMA-Update** (Exponential Moving Average) des Student-Netzwerks aktualisiert.

Dadurch lernt DINO **semantisch sinnvolle Features** zu erkennen, komplett ohne Labels oder Texte.

Steuerung durch Textprompts

Drei zentrale Mechanismen:

1. Cross-Attention

- Das Modell schaut bei jedem Schritt „auf den Prompt“.
  → Welche Wörter sind gerade relevant?

2. Conditioning

- Textvektoren modulieren interne Schichten des Modells.

3. Guidance

- Das Modell berechnet gleichzeitig:
  - Generierung **mit** Prompt
  - Generierung **ohne** Prompt
    → Differenz = *„Was will der Text?“*
    → *Verstärkung = präzisere Kontrolle*
- **Ergebnis: Textprompts sind präzise Steueranweisungen.**

Was Prompts steuern können

Prompts kontrollieren:

- **Inhalt:** Objekte, Szenen, Handlungen
- **Eigenschaften:** Größe, Farbe, Formen
- **Stil:** Fotorealismus, Cartoon, Surrealismus
- **Atmosphäre:** moody, dramatic, warm
- **Komposition:** centered, top-down, wide shot

- **Negativprompts:** „no text“, „no blur“, „no watermark“

Ein einziges Wort kann die gesamte Generierung verändern.

Prompt-Beispiele

- **„A small cat on a bed.”**
  → Größe ändert die Attribut-Dimension im Embedding
- **„A cat in front of a House.”**
  → Text kodiert räumliche Beziehungen
- **„A cat. Cozy atmosphere. Dimm lighting.”**
  → Stil- und Lichtdimensionen werden aktiviert
- Ein einziges Wort kann die gesamte Ausgabe verändern.

Generative Modelle für Text→Modalitäten

Diffusionsmodelle (z. B. Stable Diffusion)

- Entfernen schrittweise Rauschen
- Text steuert jeden Schritt durch Cross-Attention

Video-Modelle (z. B. Sora, Runway Gen-3)

- Video als Sequenz latenter Tokens
- Text sorgt für zeitliche Konsistenz

Audiomodelle (z. B. AudioLM, MusicGen)

- Audio wird Frame für Frame vorhergesagt
- Text bestimmt Stil, Instrumente, Struktur

Zusammenfassung

- Text wird in Embeddings umgewandelt, die als Steuervektor dienen
- CLIP verbindet Text & Bild
- DINO liefert starke visuelle Features
- Generative Modelle folgen dem Textvektor durch Cross-Attention, Conditioning und Guidance
- Text bestimmt Inhalt, Stil, Atmosphäre und Struktur

Quellen

- Radford et al. (2021): *Learning Transferable Visual Models From Natural Language Supervision arXiv:* https://arxiv.org/abs/2103.00020
- CLIP: Connecting Text and Images https://openai.com/index/clip/
- Computerphile – *How AI "Understands" Images (CLIP)* https://www.youtube.com/watch?v=KcSXcpluDe4
- Awesome CLIP GitHub https://github.com/yzhuoning/Awesome-CLIP
- Edge AI and Vision Alliance. (2023, January). *From DALL·E to Stable Diffusion: How do text-to-image generation models work?* https://www.edge-ai-vision.com/2023/01/from-dall%C2%B7e-to-stable-diffusion-how-do-text-to-image-generation-models-work/
- Ultralytics. (n.d.). Vision Transformer (ViT) – Erklärung und Glossareintrag. https://www.ultralytics.com/de/glossary/vision-transformer-vit
- Meta AI. (2023). *DINOv2: Advancing self-supervised learning for computer vision*. https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., et al. (2023). *DINOv2: Learning robust visual features without supervision*. arXiv. https://arxiv.org/pdf/2304.07193
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., et al. (2021). *Emerging properties in self-supervised vision transformers*. *ICCV.* https://openaccess.thecvf.com/content/ICCV2021/papers/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf