

Seminar Text Analytics

Image Understanding

**Lernen visueller Konzepte bei
Inferieren ohne fine-tuning**

**Goethe-Universität Frankfurt
Kenan Khauto**

Agenda

1. Einführung
2. Analyse von CLIP
3. Kontrastives Lernen im Detail
4. Anwendungen
5. Herausforderungen und Grenzen
6. Mögliche Verbesserungen
7. Literatur

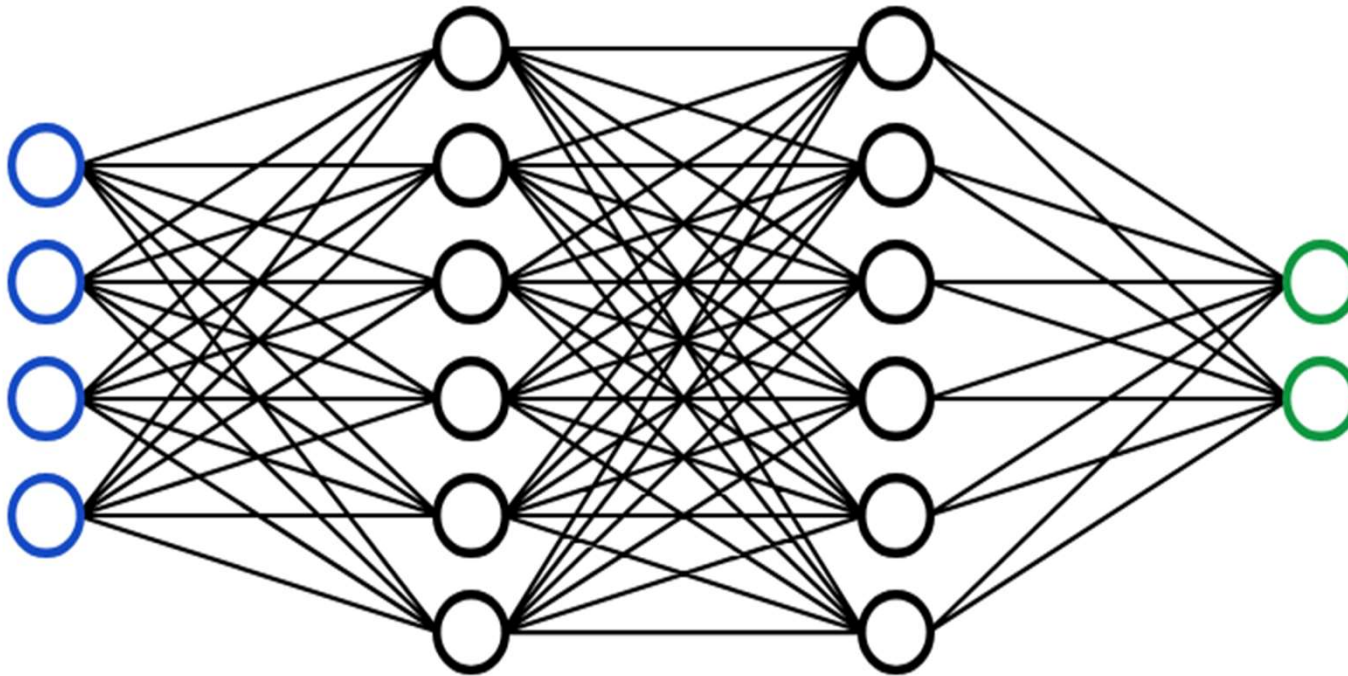
Einführung

Grundlagen des maschinellen Lernens

- Überwachtes Lernen: Lernen von Beispieldaten
- Unüberwachtes lernen: Muster oder Struktur
- Verstärkendes Lernen: durch Belohnungen lernen, um Ziel zu erreichen

Einführung

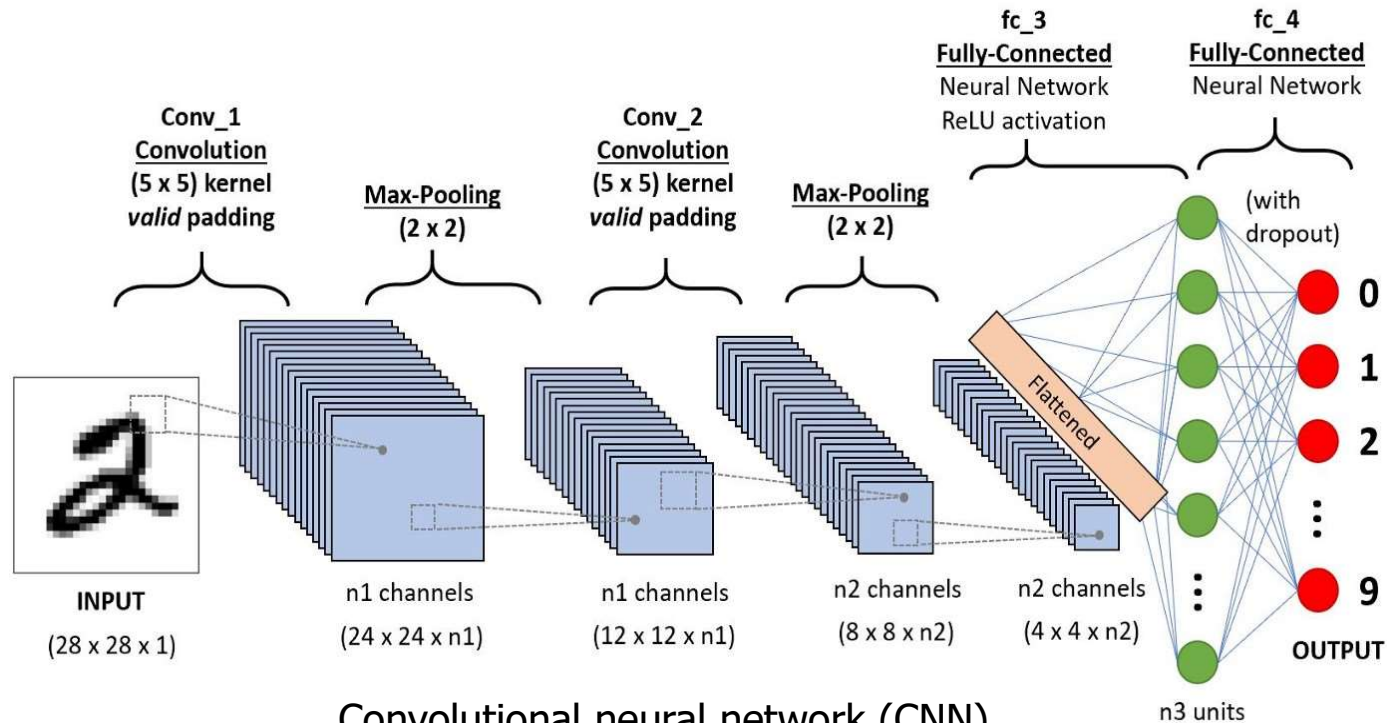
Deep Learning und neuronale Netzwerke



A simple fully connected neural network,
<https://victorzhou.com/series/neural-networks-from-scratch/>

Einführung

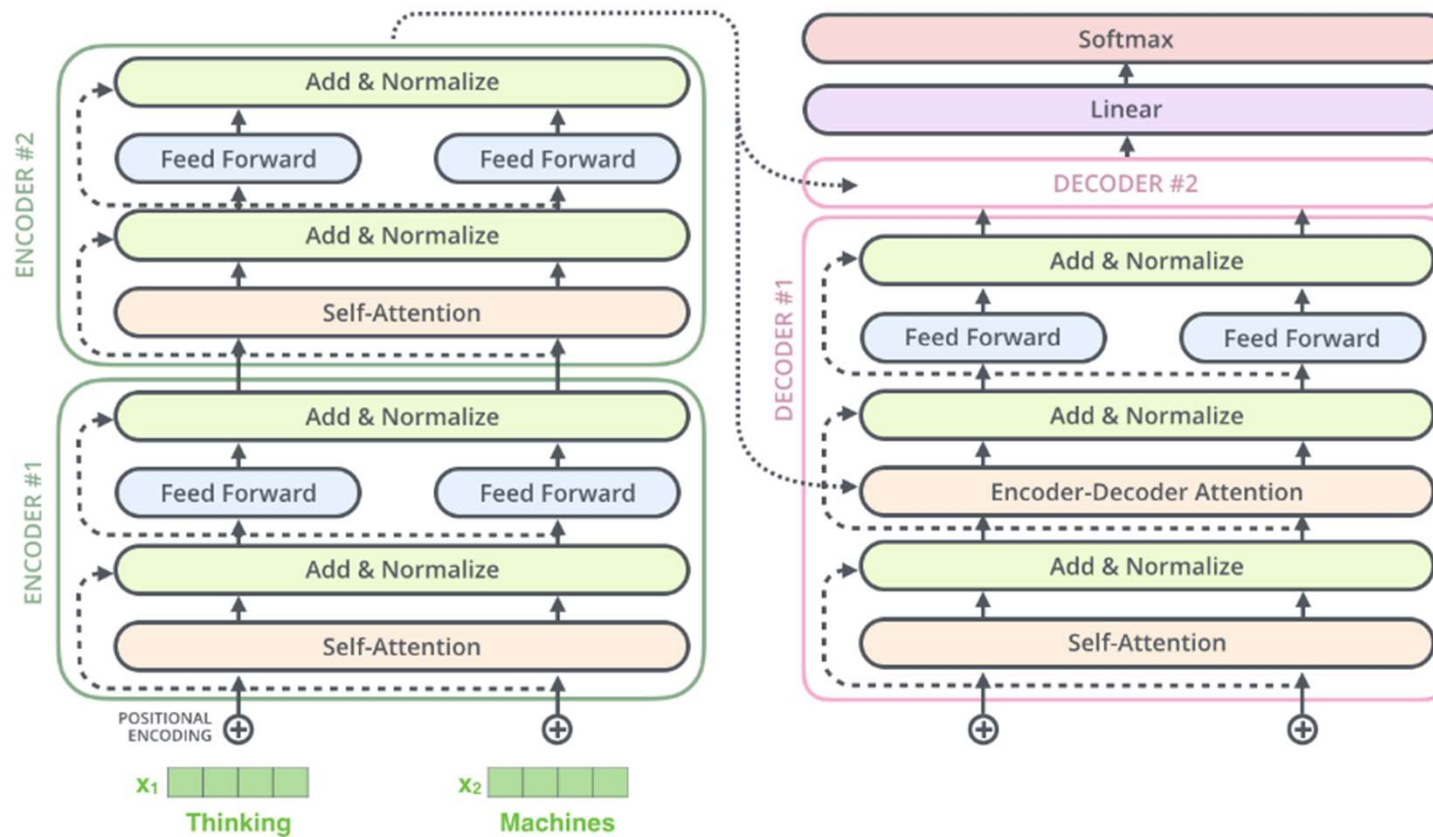
Convolutional Neural Networks (CNNs)



Convolutional neural network (CNN),
<https://paperswithcode.com/methods/category/convolutional-neural-networks>

Einführung

Transformers



Transformer mit 2 Encoders, 2 Decodern und FCL,
<https://jalammar.github.io/illustrated-transformer/>

Einführung

Kontrastives Lernen und CLIP (Radford u. a. 2021)

- **Kontrastives Lernen** ist eine Technik, die darauf abzielt, ähnliche Datenpunkte näher zusammenzubringen und unähnliche weiter von einander zu entfernen. CLIP benutzt diese Technik, um die Beziehungen zwischen Bildern und Text zu verstehen.
- **CLIP** wird mit einer Vielzahl von Bildern und den dazugehörigen Textbeschreibungen trainiert.
- **Vorteile gegenüber traditionellen Ansätzen:** CLIP kann vielfältige visuelle Konzepte anhand seiner Trainingsdaten erkennen und interpretieren.

Einführung

Seminarfrage

Wie können KI-Modelle wie CLIP visuelle Konzepte effektiv durch Inferenz verstehen und interpretieren, ohne dass ein umfangreiches Fine-Tuning erforderlich ist ?

Analyse von CLIP

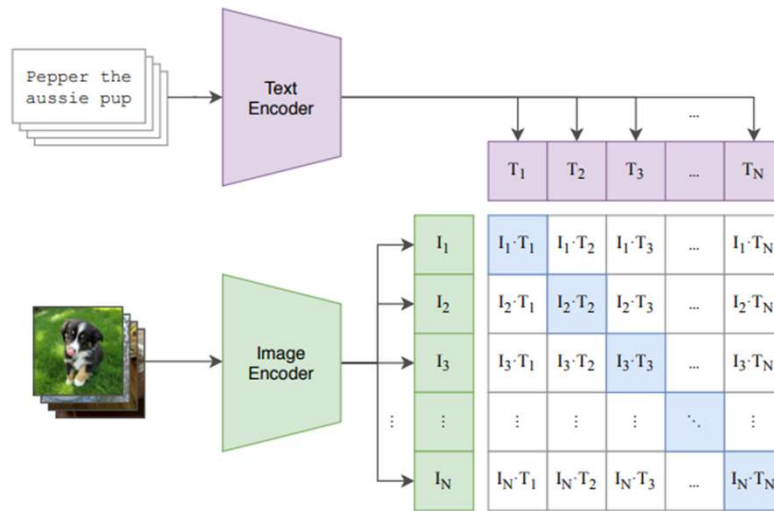
Was ist CLIP? (Radford u. a. 2021)

- CLIP ist ein neuronales Netzwerk, das anhand einer Vielzahl von Bildern und den zugehörigen Textbeschreibungen trainiert wurde.
- Dieses Training ermöglicht es ihm, sowohl visuelle als auch textuelle Eingaben zu verstehen und zu interpretieren.

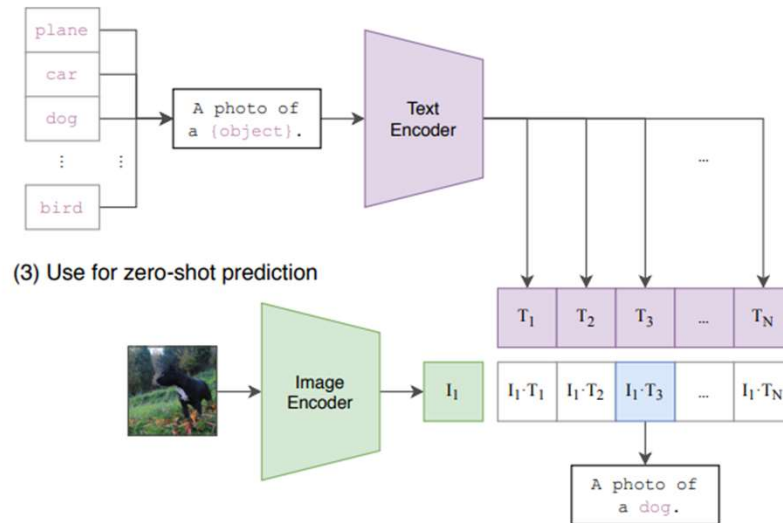
Analyse von CLIP

Architektur (Radford u. a. 2021)

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

CLIP, Bild von (Radford u. a. 2021)

Analyse von CLIP

Architektur (Radford u. a. 2021)

Das Modell besteht aus zwei Hauptkomponenten:

- **Text-Encoder:** Verarbeitet textuelle Eingaben
- **Bild-Encoder:** Verarbeitet visuelle Eingaben

Beide Encoders wandeln ihre Eingaben in einen **gemeinsamen Einbettungsraum** um, was dem Modell ermöglicht, die beiden unterschiedlichen Datentypen direkt zu vergleichen und zu verknüpfen.

Analyse von CLIP

Text-Encoder (Vaswani u. a. 2017)

- Der Text-Encoder in CLIP basiert auf einer Transformer-Architektur.
- Jedes Wort im Text wird in eine Vektordarstellung umgewandelt.
- Mit self-attention wird die Bedeutung jedes Wortes im Kontext des gesamten Text verstanden.
- **Ziel:** eine Darstellung des Textes zu erzeugen, die dessen semantischen Inhalt in Bezug auf mögliche Bildinhalte widerspiegelt.

Analyse von CLIP

Bild-Encoder (O'Shea und Nash 2015)

Der Bild-Encoder ist nicht ausschließlich auf CNN beschränkt, sondern kann auch Vision Transformers umfassen.

- CNN-basierte Encoder:
 - Traditionelle CNN-Architekturen
 - Effektiv in der Erkennung lokaler Muster
 - Wandeln das Bild in eine Vektorrepräsentation um, die die visuelle Inhalte des Bildes kodieren

Analyse von CLIP

Bild-Encoder (Dosovitskiy u. a. 2020)

Der Bild-Encoder ist nicht ausschließlich auf CNN beschränkt, sondern kann auch Vision Transformers umfassen.

- Vision Transformers:
 - Zerlegen das Bild in eine Sequenz von Patches
 - Die Patches werden ähnlich wie Wörter in einem Satz behandelt
 - Effektiv in der Erkennung sowohl lokale als auch globale Kontextinformationen aus dem Bild

Analyse von CLIP

Bild-Encoder (Dosovitskiy u. a. 2020)

Der Bild-Encoder ist nicht ausschließlich auf CNN beschränkt, sondern kann auch Vision Transformers umfassen.

Das Hauptzweck des Bild-Encoders, unabhängig von der gewählten Architektur, besteht darin, eine Darstellung des Bildes zu erzeugen, die in **demselben Vektorraum** wie der Text-Encoder liegt, um eine direkte Vergleichbarkeit zu ermöglichen.

Analyse von CLIP

Integration und gemeinsamer Einbettungsraum (Radford u. a. 2021)

Die zentrale Innovation von CLIP besteht darin, dass beide Encoder – der Text- und der Bild Encoder – darauf trainiert sind, ihre Ausgaben in einem gemeinsamen Einbettungsraum zu repräsentieren.

- Vektorraumbildung: Transformation in hochdimensionalem Vektorraum.
- Korrespondierende Paare nah beieinander
- Kontrastives Lernen: Minimierung der Distanz zwischen übereinstimmenden Bild-Text-Paaren und Maximierung der Distanz zwischen nicht übereinstimmenden Paaren
- Effektive Abbildung beider Modalitäten







Analyse von CLIP

Zero-Shot Learning in CLIP (Radford u. a. 2021)

- Zero-shot bedeutet die Anwendung eines Modells ohne die Notwendigkeit des Feintunings.
- Durch das Training mit einem vielfältigen Datensatz lernt CLIP eine breite Palette von visuellen Stilen, Objekten und Konzepten, was ihm eine gute Generalisierung auf neue, ungesehene Aufgaben ermöglicht.

Analyse von CLIP

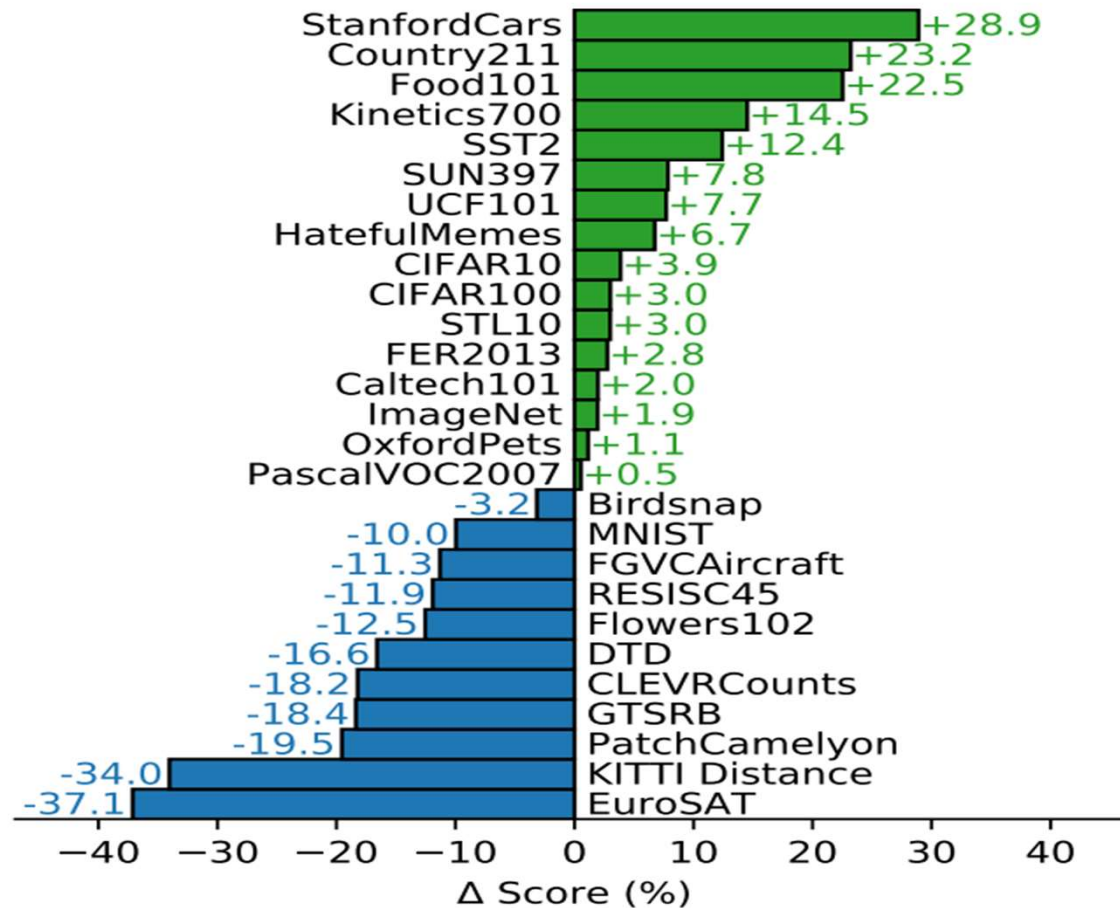
Vergleichsanalyse (Radford u. a. 2021)

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

Resnet-101 fine-tuned vs. zero-shot CLIP, Bild von (Radford u. a. 2021)

Analyse von CLIP

Vergleichsanalyse (Radford u. a. 2021)



Linear Probe ResNet50 vs. zero-shot CLIP, Bild von (Radford u. a. 2021)

Kontrastives Lernen im Detail

Vektoreinbettungen (Radford u. a. 2021)

Sei I ein Bild und T ein Text, dann

Bild-Einbettung: $v_I = f_{Bild}(I)$

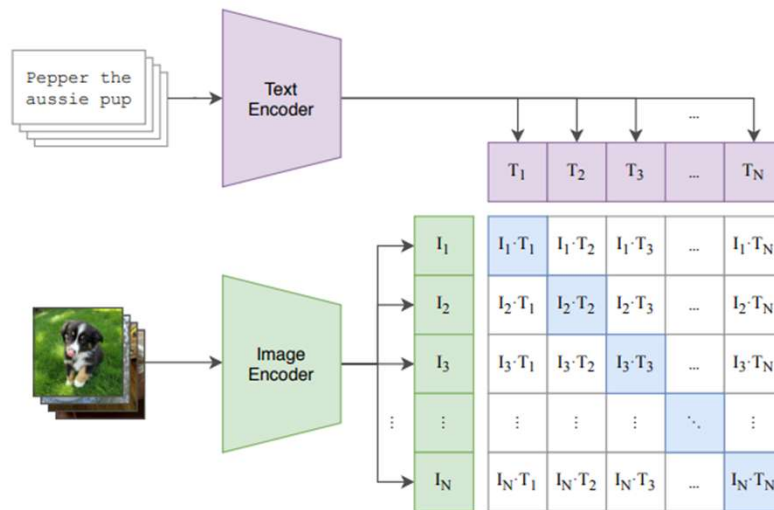
Text-Einbettung: $v_T = f_{Text}(T)$

Wobei $f_{Bild}(I)$ und $f_{Text}(T)$ die Funktionen des Bild- bzw. Text-Encoders sind.

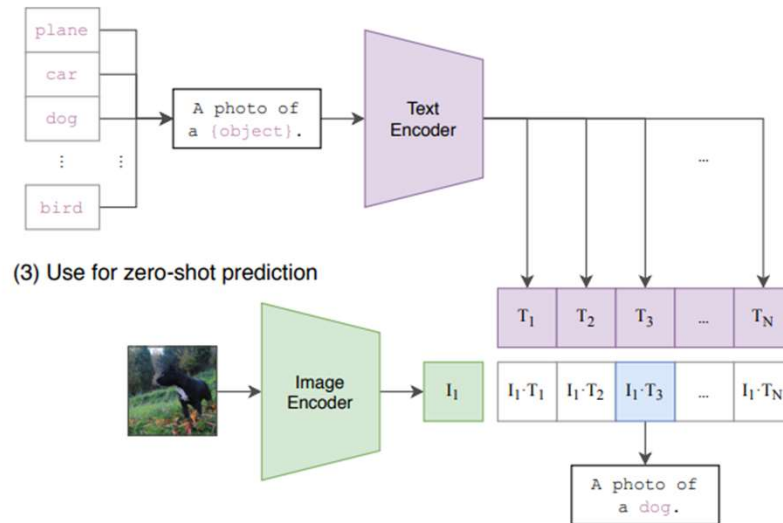
Kontrastives Lernen im Detail

Ähnlichkeitsberechnung (Radford u. a. 2021)

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

CLIP, Bild von (Radford u. a. 2021)

Kontrastives Lernen im Detail

Ähnlichkeitsberechnung (Radford u. a. 2021)

Die Ähnlichkeit zwischen einem Bild- und einem Textvektor wird durch das Skalarprodukt ihrer normalisierten Vektoren berechnet:

$$\text{sim}(\mathbf{v}_I, \mathbf{v}_T) = \frac{\mathbf{v}_I \cdot \mathbf{v}_T}{\|\mathbf{v}_I\| \cdot \|\mathbf{v}_T\|}$$

Dann wird die Verlust für ein Paar (\mathbf{I}, \mathbf{T}) berechnet als:

$$\mathcal{L}(\mathbf{I}, \mathbf{T}) = -\log \frac{\exp\left(\frac{\text{sim}(\mathbf{v}_I, \mathbf{v}_T)}{\tau}\right)}{\sum_{T'} \exp\left(\frac{\text{sim}(\mathbf{v}_I, \mathbf{v}_{T'})}{\tau}\right)} - \log \frac{\exp\left(\frac{\text{sim}(\mathbf{v}_T, \mathbf{v}_I)}{\tau}\right)}{\sum_{I'} \exp\left(\frac{\text{sim}(\mathbf{v}_T, \mathbf{v}_{I'})}{\tau}\right)}$$

- τ ist ein Temperatur-Parameter
- Die Summen laufen über alle Texte und Bilder jeweils

Kontrastives Lernen im Detail

Gesamtverlust (Radford u. a. 2021)

Für ein Batch mit N Bild-Text Paare, wird die Gesamtverlust mit:

$$L_{total} = \frac{1}{N} \sum_{i=1}^N L(I_i, T_i)$$

berechnet.

Die symmetrische Natur der Loss-Funktion (symmetric Cross Entropy Loss) stellt sicher, dass die Bild-zu-Text- und Text-zu-Bild-Vorhersagen während des Trainings gleichmäßig betont werden

Anwendungen

Anwendungen

- **Zero-Shot-Lernen:** CLIP kann Bilder klassifizieren, in dem es die Ähnlichkeiten zwischen Bildern und Textbeschreibungen nutzt.
- **Bild- und Textsuche:** Es ermöglicht die Suche nach Bildern mit textuellen Beschreibungen und umgekehrt, die Suche nach Texten, die zu einem Bild passen.
- **Content Moderation:** CLIP kann zur automatischen Erkennung und Filterung unangemessener Inhalte in Bildern und Texten eingesetzt werden.
- **Automatische Bildbeschriftung:** Generierung von Textbeschreibungen für Bilder.
- **Sentiment-Analysis in Bildern:** Erkennung der Stimmung oder des Gefühls, das ein Bild vermittelt.
- Und noch mehr ...

Anwendungen

Beispielcode für Bild-Klassifizierung (Pinecone 2023)

```
from tqdm.auto import tqdm

preds = []
batch_size = 32

for i in tqdm(range(0, len(imagenette), batch_size)):
    i_end = min(i + batch_size, len(imagenette))
    images = processor(
        text=None,
        images=imagenette[i:i_end]['image'],
        return_tensors='pt'
    )['pixel_values'].to(device)
    img_emb = model.get_image_features(images)
    img_emb = img_emb.detach().cpu().numpy()
    scores = np.dot(img_emb, label_emb.T)
    preds.extend(np.argmax(scores, axis=1))
```

Der komplette Code findet man auf
<https://github.com/KenanKhauto/zero-shot-learning>

```
true_preds = []
for i, label in enumerate(imagenette['label']):
    if label == preds[i]:
        true_preds.append(1)
    else:
        true_preds.append(0)

sum(true_preds) / len(true_preds)
```

✓ 0.0s

0.9831847133757962

Grenzen

Grenzen

- **Verzerrungen und Vorurteile:** Wie viele KI-Modelle kann auch CLIP-Verzerrungen aufweisen, die in den Trainingsdaten vorhanden sind. Dies kann zu unfairen oder voreingenommenen Ergebnissen führen, besonders bei der Analyse von Bildern und Texten aus verschiedenen Kulturen und sozialen Gruppen.
- **Abhängigkeit von der Textqualität:** Die Leistung von CLIP ist stark abhängig von der Qualität und Relevanz der Textbeschreibungen. Unpräzise oder irreführende Texte können zu fehlerhaften Ergebnissen führen.
- **Generalisierungsfähigkeit:** Obwohl CLIP gut in der Lage ist, Konzepte zu generalisieren, kann es Schwierigkeiten geben, sehr spezifische oder seltene Objekte und Szenarien korrekt zu erkennen und zuzuordnen.
- **Komplexität und Ressourcenanforderungen:** CLIP-Modelle sind groß und rechenintensiv, was ihre Anwendbarkeit in ressourcenbeschränkten Umgebungen einschränkt.

Mögliche Verbesserungen

Verbesserungen

- **Diversifizierung der Trainingsdaten:** Um Verzerrungen und Vorurteile zu reduzieren, sollten die Trainingsdaten vielfältiger und repräsentativer gestaltet werden. Dies schließt Daten aus verschiedenen Kulturen, Sprachen und sozialen Gruppen ein.
- **Erweiterte Kontextanalyse:** Die Integration zusätzlicher Kontextinformationen kann helfen, die Genauigkeit der Bild-Text-Zuordnungen zu verbessern, besonders bei komplexen oder mehrdeutigen Szenen.
- **Interdisziplinäre Ansätze:** Zusammenarbeit mit Experten aus verschiedenen Bereichen wie Sozialwissenschaften, Ethik und Kunst, um die Anwendungen und Implikationen von CLIP besser zu verstehen und zu steuern.

Quellen

- Dosovitskiy, Alexey u. a. (2020). „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“. In: *arXiv preprint arXiv:2010.11929*.
- O'Shea, Keiron und Ryan Nash (2015). „An Introduction to Convolutional Neural Networks“. In: *arXiv preprint arXiv:1511.08458*.
- Pinecone (2022). *Zero Shot Object Detection with OpenAI's CLIP*. Accessed on: 2024-01-22. URL: <https://www.pinecone.io/learn/series/image-search/zero-shot-object-detection-clip/>.
- (2023). *Zero-shot Image Classification with OpenAI's CLIP*. Accessed on: 2024-01-22. URL: <https://www.pinecone.io/learn/series/image-search/zero-shot-image-classification-clip/>.
- Radford, Alec u. a. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv: [2103.00020](https://arxiv.org/abs/2103.00020) [cs.CV].
- Vaswani, Ashish u. a. (2017). „Attention is all you need“. In: *Advances in neural information processing systems*, S. 5998–6008.

Fragen ?