

M09 - Multimodal Bild

Stand: 03.2025

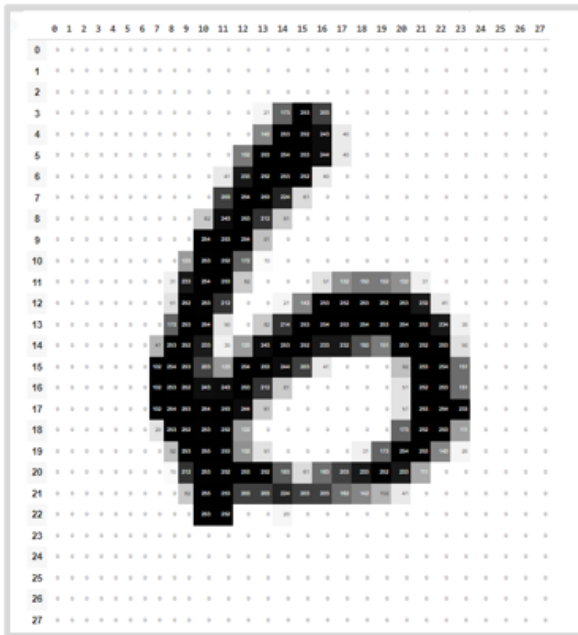
1 Grundlagen Bilderkennung

Ein Computer "sieht" ein Bild nicht wie ein Mensch. Für ihn stellt ein Bild lediglich eine Matrix aus Zahlenwerten dar, wobei jeder Wert die Intensität eines Pixels in verschiedenen Farbkanälen (meist Rot, Grün, Blau) repräsentiert. Diese Zahlenwerte werden von Algorithmen verarbeitet, um Muster und Strukturen zu erkennen, die für die Bilderkennung essenziell sind.

Der Prozess der Bilderkennung umfasst folgende Schritte:

1. **Bilderfassung:** Ein Bild wird in eine numerische Darstellung umgewandelt. Dabei werden Bildformate wie JPEG oder PNG in ein Raster aus Pixelwerten umgerechnet.
2. **Vorverarbeitung:** Das Bild wird normalisiert, skaliert und aufbereitet. Dies kann Schritte wie Rauschunterdrückung, Kontrastanpassung oder Farbnormalisierung umfassen.
3. **Merkmalsextraktion:** Wichtige Merkmale wie Kanten, Formen oder Farbverteilungen werden identifiziert. Hier können Methoden wie Histogramm-basierte Ansätze oder Kantendetektion (z. B. Sobel-Operator) angewendet werden.
4. **Klassifikation:** Basierend auf den extrahierten Merkmalen erfolgt eine Klassifikation des Bildes. Dies geschieht mithilfe von maschinellen Lernmodellen oder regelbasierten Systemen.

Raster aus Pixelwerten



2 Methoden

2.1 Traditionelle Methoden

Bei traditionellen Methoden müssen explizit Merkmale definiert werden, die als relevant gelten (z. B. Kanten, Farben, Texturen). Klassische Verfahren beinhalten Methoden wie:

- **Kantendetektion** mittels Sobel- oder Canny-Operator
- **Merkmalsvektoren** wie SIFT (Scale-Invariant Feature Transform) oder HOG (Histogram of Oriented Gradients)
- **Template Matching**, um spezifische Muster in Bildern zu finden

Diese Verfahren erfordern umfassendes domänenspezifisches Wissen und sind oft anfällig für Variationen in Beleuchtung, Perspektive oder Bildrauschen.

Merkmals-Filter: <https://editor.p5js.org/ralf.bendig.rb/full/zLXqi5u6f>

Merkmals-Filter-Anwendung: <https://editor.p5js.org/ralf.bendig.rb/full/Xi2uabjR9>

2.2 Deep Learning

Moderne Ansätze setzen auf neuronale Netze, insbesondere Convolutional Neural Networks (CNNs), die eigenständig lernen, welche Merkmale relevant sind. CNNs bestehen aus mehreren Schichten, die folgende Aufgaben erfüllen:

- **Faltungsschichten (Convolutional Layers)**: Extrahieren Merkmale durch das Anwenden von Filtern
- **Pooling-Schichten**: Reduzieren die Dimensionen und verallgemeinern die Merkmale

- **Voll verbundene Schichten (Fully Connected Layers):** Nutzen die extrahierten Merkmale zur Klassifikation

Deep-Learning-Modelle werden auf große Datensätze trainiert, wodurch sie eine hohe Generalisierungsfähigkeit erreichen und in der Lage sind, komplexe Muster autonom zu lernen.

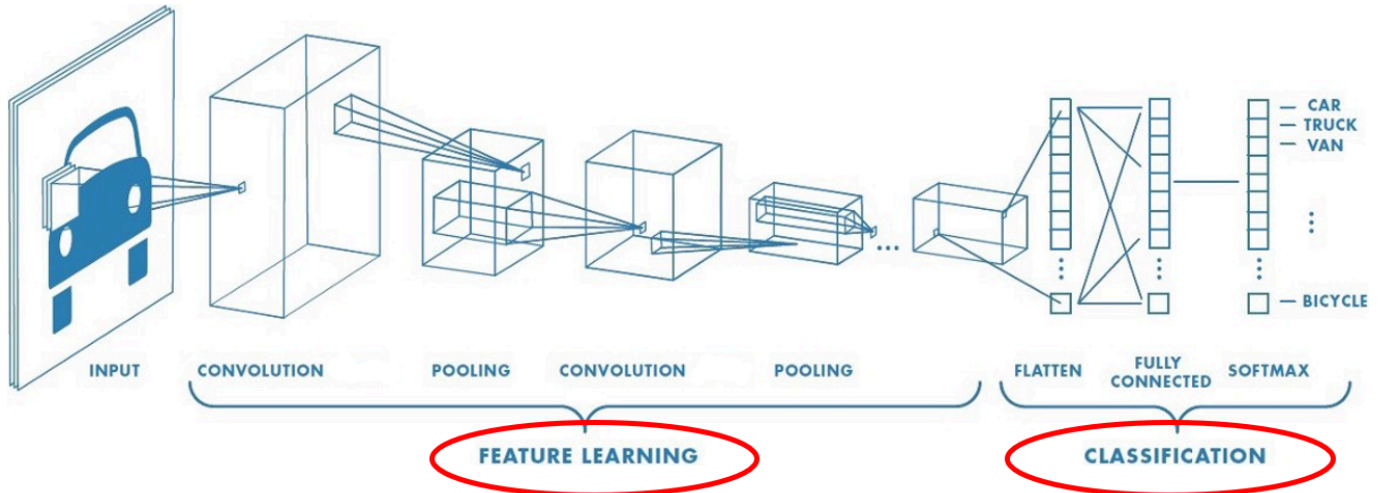


Bild: [A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science](#)

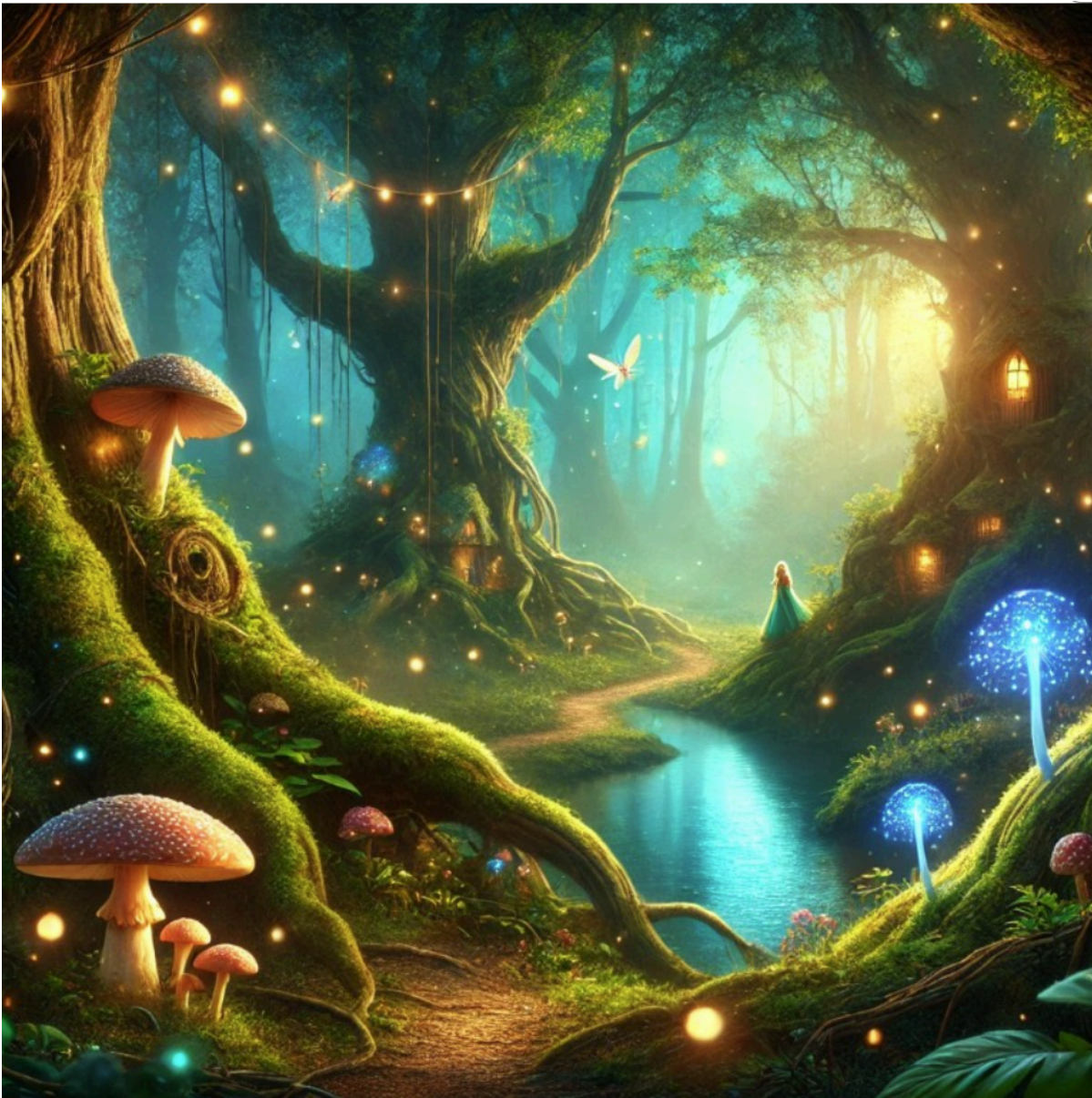
3 Bild-Modelle

3.1 Text-zu-Bild-Modelle

In diesem Abschnitt wird der Fokus auf Text-zu-Bild-Modelle gelegt, einem innovativen Bereich der künstlichen Intelligenz, der es Maschinen ermöglicht, Bilder basierend auf Textbeschreibungen zu erzeugen. Diese Modelle überbrücken die Kluft zwischen Sprache und visuellen Inhalten, indem sie eine natürliche Spracheingabe in eine vollständige Bilddarstellung umsetzen. Die Generierung von Bildern aus Text basiert auf Deep-Learning-Methoden, insbesondere durch die Kombination von natürlicher Sprachverarbeitung (NLP) und Computer Vision.

Ein zentrales Merkmal dieser Modelle ist ihre Fähigkeit, Zusammenhänge zwischen sprachlichen und visuellen Elementen zu erfassen. Durch das Training mit umfangreichen Datensätzen, die Texte mit den dazugehörigen Bildern verknüpfen, lernen sie, sprachliche Beschreibungen wie „eine Katze sitzt auf einem Fensterbrett“ mit relevanten visuellen Eigenschaften wie Formen, Texturen, Farben und räumlichen Anordnungen zu verbinden. Modelle wie DALL·E, MidJourney und Stable Diffusion haben das kreative Potenzial dieser Technologie eindrucksvoll demonstriert, indem sie sowohl fotorealistische als auch künstlerische Bilder direkt aus textlichen Vorgaben generiert haben.

Hier wird veranschaulicht, wie DALL·E ein Bild mit einem zauberhaften Märchenwald generiert:



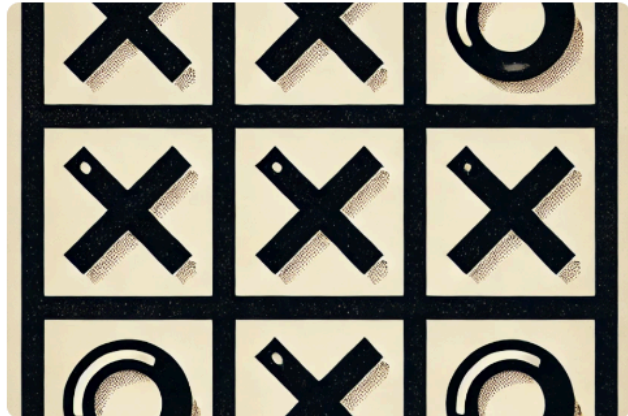
3.2 Multimodale Modelle

In diesem Abschnitt wird die spannende Welt multimodaler Modelle untersucht, die verschiedene Datentypen wie Text, Bilder, Audio und Video verarbeiten und zu einem ganzheitlichen Verständnis verknüpfen. Diese Modelle markieren einen bedeutenden Fortschritt in der KI, da sie Eingaben aus unterschiedlichen Quellen kombinieren und dadurch komplexere Aufgaben bewältigen können – ähnlich wie der Mensch Informationen aus verschiedenen Sinneseindrücken zusammenführt.

Während spezialisierte Modelle, etwa Text-zu-Bild-Systeme, bereits einzelne Verbindungen zwischen Modalitäten ermöglichen, gehen multimodale Modelle einen Schritt weiter. Sie analysieren und verstehen die Beziehungen zwischen verschiedenen Informationsarten, was vielfältige Anwendungen ermöglicht. Dazu zählen das Generieren von Bildunterschriften, das Beantworten visueller Fragen oder die Erstellung von Bildern auf Basis von Text und anderen Medieneingaben.

Ein anschauliches Beispiel für multimodale Fähigkeiten ist die Analyse eines handgezeichneten Tic-Tac-Toe-Bretts. Ein solches Modell kann ein Bild des Spiels

interpretieren, die Platzierung von X und O erkennen und basierend auf den Spielregeln den Gewinner bestimmen – ohne zusätzliche textliche Informationen über das Spielfeld zu benötigen.



Wer hat das Spiel gewonnen?

Das Bild zeigt ein Tic-Tac-Toe-Spielbrett mit einem 3x3-Raster. Die meisten Felder sind mit "X" markiert, während drei Felder ein "O" enthalten.

Um den Gewinner zu bestimmen, prüfen wir, ob eine der Parteien drei ihrer Symbole in einer Reihe, Spalte oder Diagonale angeordnet hat:

- Horizontale Reihen: Alle Reihen enthalten gemischte Symbole oder nur "X".
- Vertikale Spalten: Keine Spalte enthält ausschließlich "O".
- Diagonale Linien: Keine vollständige Diagonale besteht nur aus "O".

Da in keiner Reihe, Spalte oder Diagonale drei "O" erscheinen, aber eine horizontale Reihe oder Spalte vollständig mit "X" belegt ist, hat "X" das Spiel gewonnen.



Durch die Kombination verschiedener Datentypen ermöglichen multimodale Modelle eine fortschrittlichere und intuitivere Interaktion zwischen Mensch und KI. Dies macht sie zu einem vielseitigen Werkzeug mit breiten Anwendungsmöglichkeiten – von Gesundheitswesen und Bildung bis hin zu Unterhaltung und Design.

4 Image-Embeddings

Ähnlich wie bei Text-Embeddings, die Wörter oder Sätze in einer Weise kodieren, dass semantische Ähnlichkeiten erhalten bleiben, transformieren Image-Embeddings visuelle Merkmale in eine für Maschinen lernbare Form.

Mithilfe neuronaler Netze – typischerweise Convolutional Neural Networks (CNNs) oder Transformer-Modelle wie CLIP – werden hochdimensionale Bilddaten in kompakte Vektoren

umgewandelt. Diese Embeddings ermöglichen Aufgaben wie Bildähnlichkeitssuche, Clustering oder die Kombination von Bild- und Textdaten für multimodale Modelle.

[Image-Embeddings](#)