Multimodal Computational Methods in Political Science

Images and Convolutional Neural Networks

Clara Fochler

Ludwig-Maximilians-Universität München

June 02, 2026

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Agenda

Why go beyond computational text analysis as a political scientist?

Introduction to Images

Recap: Neural Networks

Convolutional Neural Networks (CNN)

The Issue with Deeper CNNs and ResNet

Region-based CNNs for Object Detection

Image Preprocessing and CNN Autotaggers

Key Takeaways and What's Next

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Why go beyond computational text analysis?

- increased use of images with digital space - many mostly used social media sites nowadays rely heavily on image, audio and video - including memes and AI-images

As Webb, Casas and Wilkerson (2020) summarizes previous research -including images in social science analysis is necessary because:

- Images capture attention more effectively than written or spoken content
- Visuals impact agenda-setting, political participation, perception of politicians
- They facilitate faster information processing and improve memory recall
- Visuals are uniquely capable of evoking strong emotional responses

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

From Text to Images: Same Logic, Different Data

Concept	Text Analysis	Image Analysis
Basic Unit	Token (Word/Subword)	Pixel
Input Structure	Sequence of Tokens	Grid of Pixels (H $\times$ W $\times$ 3)
Representation	Word Embedding	Feature Map
Core Operation	Attention over Tokens	Convolution over Patches
Pretrained Model	BERT, RoBERTa	ResNet, VGG
Adaptation	Fine-tune on labeled text	Fine-tune on labeled images

- **Political text:** *"We stand together against the far right!"* $\rightarrow$ Tokenization $\rightarrow$ Sentiment / Frame Classification
- **Political image:** Protest photo with banners $\rightarrow$ Pixel grid $\rightarrow$ Object Detection / Scene Classification

–> The core workflow is identical - but images capture what text cannot: bodies, symbols, scale, and emotion.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Images (1/3)

- in computers images are represented by matrices or graphs displaying pixels
- file-types: JPEG, PNG, RAW, TIF, etc.
- quality of digitized image is dependent on sampling (number of pixels) and quantization (color depth)
- The largest and most often used labeled Image Dataset for CNN Pretraining is ImageNet; MS-Celeb-1M contains celebrity labeled images; COCO contains images labeled by objects
- Additional Datasets can be found on Kaggle or Huggingface or Harvard Dataverse
- In social sciences we mostly use computer vision for: classification, object detection and sentiment analysis

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Images (2/3)

Images can be characterized by:

- Spatial dimensions: 2D (traditional images) vs. 3D (volumetric data like satellite/drone imagery)
- Color system: RGB, grayscale, HSV, CMYK, or other color spaces
- Temporal component: Static (single frame/image) vs. dynamic (video/sequence of frames over time)

Introduction to Images (3/3)

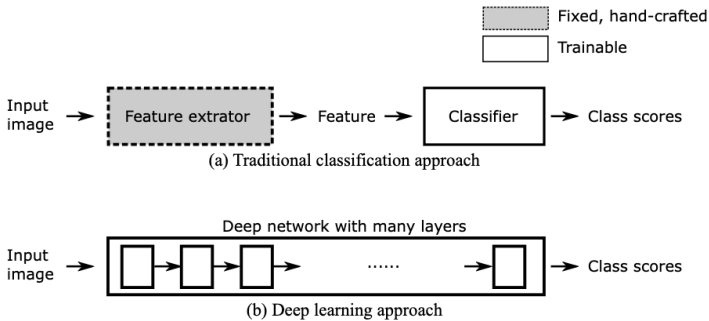


Figure 6: Comparing Deep Learning to Previous Computer Vision Methods

Figure: Taken from Joo and Steinert-Threlkeld (2018)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Recap: Neural Networks

You have already learned in Section 1 of this seminar, what a Neural Network is.

- Inspired by neurons in brains (artifical neurons are also called "perceptron")
- Neurons calculate weighted sum of input values
- A neural network learns by passing data forward through weighted layers (**forward pass**), computing the error via a **loss function**, and updating the weights backwards through the network (**backpropagation**) using **gradient descent**
- In a Dense Neural Networks - each layer of Neurons receives the same inputs –> when analyzing images, Dense Neural Networks must flatten e.g. 2D image to 1D vector and each neuron gets weights for each input - meaning that when you have 1000 inputs and 10 neurons you will get 10 000 weights
- When more than one hidden layer –> called "Deep Neural Network"

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CNNs versus Dense NNs

Why CNNs for Image Analysis instead of NNs?

- For Dense NNs each input neuron multiplies with a different weight for each output neuron
- For Convolutional NNs the same kernel multiplies with many different patches of neurons, at different locations –> local structure matters

Convolution

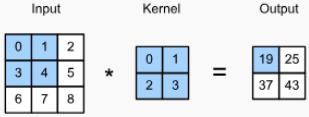


Fig. 7.2.1 Two-dimensional cross-correlation operation. The shaded portions are the first output element as well as the input and kernel tensor elements used for the output computation: $0 \times 0 + 1 \times 1 + 3 \times 2 + 4 \times 3 = 19$.

Figure: Taken from Zhang et al. (2024)

- **Tidbit:** In most CNN libraries (e.g. PyTorch) we actually use Cross-Correlation instead of Convolution (Convolution would require the kernel to be flipped - since mathematically the order of the kernel should not change the result for convolution)

Convolutional Neural Networks (CNN)

CNNs rely on three core principles:

- Translation Invariance: Early layers recognize the same pattern anywhere in the image.
- Locality: They focus on small, local areas first.
- Hierarchy: Deeper layers aggregate these local details to understand the whole image. –> Receptive Field refers to the fields neurons in each layer can see - meaning the deeper the CNN, the larger the receptive field for each added neuron layer
  –> this leads to the intuition, that the deeper the CNN, the better it can represent complex features

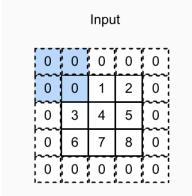
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Images as 3rd-order Tensors

- CNNs process data as 3rd-order tensors (*Height* $\times$ *Width* $\times$ *Channels*) - not flat images.
- The **Channel** dimension means different things at different stages:
  - **Input:** 3 fixed channels: Red, Green, Blue (RGB).
  - **Hidden Layers:** *n* learned channels, each encoding a detected feature (e.g., edges, textures, shapes).
- In other words: depth evolves from **raw color** $\rightarrow$ **abstract features** as data passes through the network.

Padding, Stride and Pooling

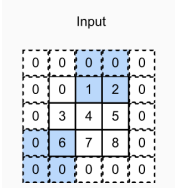
Padding



Taken from Zhang et al. (2024)

- Adds extra pixels (zeros) around input borders
- Preserves spatial dimensions of output
- Prevents edge information loss
- Enables deeper network architectures

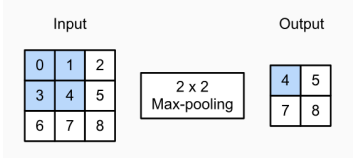
Stride



Taken from Zhang et al. (2024)

- Steps a kernel shifts across the input per move
- Stride $S{=}1$: one element at a time
- Stride $S{>}1$: skips S elements per step
- Used for downsampling & efficiency

Pooling



Taken from Zhang et al. (2024)

- Reduces spatial dimensions & locality sensitivity
- **Max pooling**: takes maximum in filter window (common)
- **Avg pooling**: takes arithmetic mean
- Custom stride & padding also applicable

The Issue with Deeper CNNs

Deeper models initially improved performance (AlexNet, VGG), but eventually depth became a liability: training error increased and accuracy saturated, suggesting a fundamental challenge with very deep networks.

The Problem

- Neural networks should improve when we add more layers, but empirically they often degrade
- Deeper networks don't simply extend what shallower networks learned - they represent fundamentally different function classes
- A deeper architecture may learn patterns that move away from the target function rather than toward it
- No guarantee that larger model capacity translates to better solutions

Why This Happens

- Expanding from a shallow to a deep network changes what the model can express
- The optimization landscape becomes more complex; the network can get stuck in worse local optima
- The larger function class doesn't necessarily contain the solutions found by the smaller class

ResNet

The ResNet Solution: Skip Connections

- Skip connections: direct pathways that allow information to bypass intermediate layers
- Each residual block learns: output = input + f(input), where f is the new layers
- If f learns to be zero (do nothing), the output is just the input; if f learns something useful, it gets added on top
- New layers can either contribute useful transformations or simply pass information through (learning the identity mapping)
- Result: adding layers is monotonically non-decreasing in performance-depth becomes a free parameter to scale safely

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CNN Autotaggers

- Pretrained CNNs (or others), that will generate image labels - possible bias issues
- Commercial (e.g. Amazon Rekognition, Microsoft Computer Vision, Google Cloud Vision) vs open-source (e.g. YOLOv8 (object detection), face_recognition)
- While open-source Autotaggers might provide fewer labels and be trained for fewer tasks, but they can be adapted via e.g. fine-tuning - which we will discuss next week
- For commercial Autotaggers we do not know the existing labels until trying and we do not know the exact model and updates they use (issue of reproducibility)
- **But** many labels we are interested in are not currently covered by autotaggers **and** might not be accurate for our tasks

Figure 6.1 Rekognition labels with confidence: Human 99.1 Bar Counter 98.4 Pub 98.4 Diner 97.1 Food 97.1 Worker 83.4 Animal 76.4 Aquarium 76.4 Sea Life 76.4.

Source: https://twitter.com/AmericaNewsroom/status/910493241283358720

Figure: Taken from Webb, Casas and Wilkerson 2020

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Region-based CNNs for Object Detection

- **R-CNN:** Propose regions, extract features per region, classify each with a CNN
- **Fast R-CNN:** Extracts features once from the full image; trains classification and localization jointly in a single end-to-end model
- **Faster R-CNN:** Introduces a Region Proposal Network (RPN), a convolutional mechanism that selectively proposes regions of interest, combined with Fast R-CNN
- **YOLO:** Divides the image into a grid and predicts bounding boxes and class labels for each cell simultaneously in a single forward pass

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Object Detection: Two-Stage vs. Single-Stage

Two-Stage (e.g., Faster R-CNN)

- *Process:* Region Proposal $\rightarrow$ Classification
- **Pros:** High accuracy, precise localization
- **Cons:** Slower due to multi-stage pipeline

Single-Stage (e.g., YOLO)

- *Process:* Direct regression in one pass
- **Pros:** Extremely fast (Real-time)
- **Cons:** Challenging with small/dense objects

In our practical session we will use YOLOv8.
–> keeps the single-pass philosophy of the original YOLO but uses a much deeper, more sophisticated convolutional backbone and detects objects at multiple scales simultaneously –> far more accurate on complex real-world images

Legal & Ethical Concerns

Copyright (EU)

- No general fair use principle in EU law.
- DSM Directive allows data mining exemption for research - but under strict conditions

Privacy

- Images often contain personal data (faces, locations)
- **GDPR:** Legal basis required for collection & processing; anonymization must be considered

Harm & Algorithmic Bias

- Non-representative training data reproduces societal biases
- We must identify bias and anticipate harm before deployment

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Image Preprocessing

- Make images the same size (padding, cropping, squishing, rescaling)
- Normalization (scaling and centering) color intensity
- If beneficial do data augmentation - can decrease necessary training set size (e.g. random crop, horizontal flip)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

REFERENCES

- Joo, J., & Steinert-Threlkeld, Z. C. (2018). Image as data: Automated visual content analysis for political science. *arXiv preprint arXiv:1810.01544*. https://arxiv.org/abs/1810.01544
- Webb Williams N, Casas A, Wilkerson JD. Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification. Cambridge University Press; 2020.
- Zhang, A., Lipton, Z. C., Li, M., Smola, A. J. (2024). Dive into deep learning. Cambridge University Press. https://d2l.ai

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thanks for your attention!