Deep Learning: More Than Classification

Calvin Seward

JOHANNES KEPLER UNIVERSITÄT LINZ

zalando research

4 September 2017

AI Summit Vienna

OVERVIEW

~~SHAMELESS SALES PITCH~~ COMPANY INFORMATION

- Europe's leading online fashion platform
- Operating in 15 countries

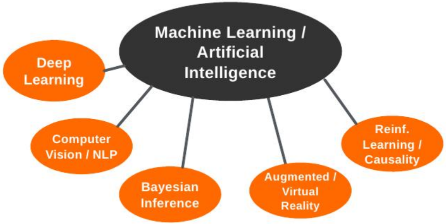
- €3.6 billion net sales 2016
- $\sim$13,000 employees from 100+ countries
- $\sim$1,800 employees in technology

- $\sim$250,000 fashion items offered
- $\sim$21 million active customers

Interested? Visit https://tech.zalando.de and https://jobs.zalando.de

~~SHAMELESS SALES PITCH~~ ZALANDO RESEARCH

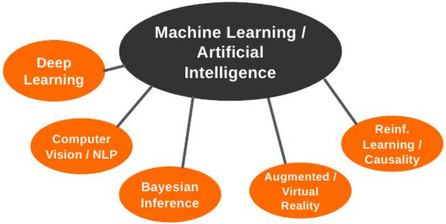
- Understand fashion & style
- Revolutionize online shopping experience
- Product impact, papers, patents, prestige
- Autonomous researchers
- Big Data, NVIDIA GPUs, Tensorflow. . .



Contact: research@zalando.de

~~SHAMELESS SALES PITCH~~ ZALANDO RESEARCH

- Understand fashion & style
- Revolutionize online shopping experience
- Product impact, papers, patents, prestige
- Autonomous researchers
- Big Data, NVIDIA GPUs, Tensorflow. . .
- We're hiring!!!



Contact: research@zalando.de

Why Neural Networks Are Powerful

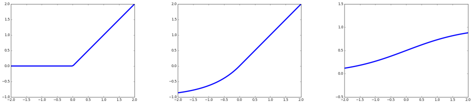
BASIC FEED FORWARD NEURAL NETWORKS

- In mathematical notation:

$$h = \sigma(W_1 \text{input} + b_1)$$
$$\text{output} = W_2 h + b_2$$

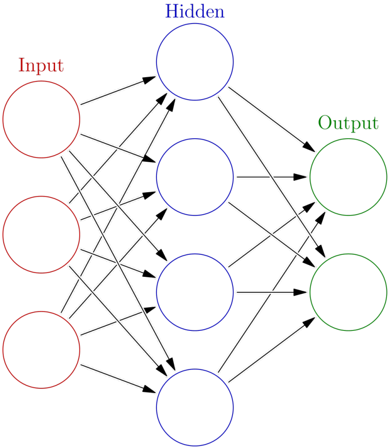
$$\hat{f}_W(\text{input}) = W_2 \sigma(W_1 \text{input} + b_1) + b_2$$

- σ, the **activation function** is a non-linear function such as



(a) ReLU (b) eLU (c) Sigmoid

Activation Functions



Source: Wikipedia

MACHINE LEARNING AS LEARNING FUNCTIONS

Most machine learning problems boil down to estimating a function $f : \mathbb{R}^n \to \mathbb{R}^m$ based on observations $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$. Some examples include:

- Image Classification: x_i is an image, $f(x_i)$ is its label
- Self Driving Cars: x_i is a video sequence, $f(x_i)$ is the next action the car should take
- Recommender Systems: x_i is a customer's shopping history, $f(x_i)$ is a product recommendation

MACHINE LEARNING AS LEARNING FUNCTIONS

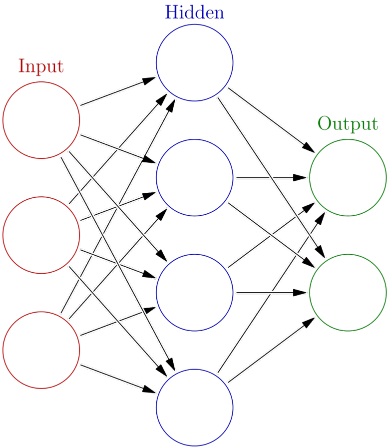
Most machine learning problems boil down to estimating a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ based on observations $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$. Some examples include:

- Image Classification: x_i is an image, $f(x_i)$ is its label
- Self Driving Cars: x_i is a video sequence, $f(x_i)$ is the next action the car should take
- Recommender Systems: x_i is a customer's shopping history, $f(x_i)$ is a product recommendation

The true function $f(x)$ will never be know, we try to find an estimate $\hat{f}$

BASIC FEED FORWARD NEURAL NETWORKS

- Neural Networks are universal function approximation
- For any function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ fulfilling certain smoothness criteria there exists a neural network $\hat{f}_W$ that is arbitrarily close to f

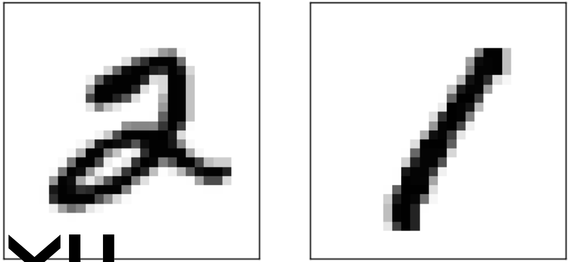


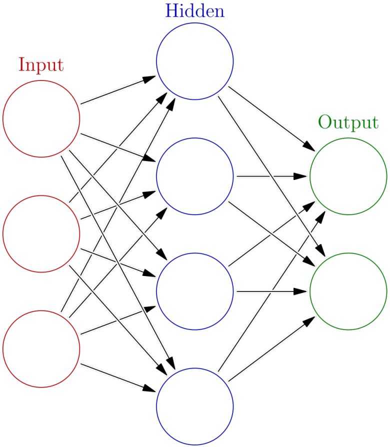
Source: Wikipedia

BASIC FEED FORWARD NEURAL NETWORKS

- Neural Networks are universal function approximation
- For any function $f : \mathbb{R}^n \to \mathbb{R}^m$ fulfilling certain smoothness criteria there exists a neural network $\hat{f}_W$ that is arbitrarily close to f
- Finding correct weights W would give us

$$\hat{f}_W : \text{Image} \to \text{Label}$$





Source: Wikipedia

BASIC FEED FORWARD NEURAL NETWORKS – BACKPROPAGATION

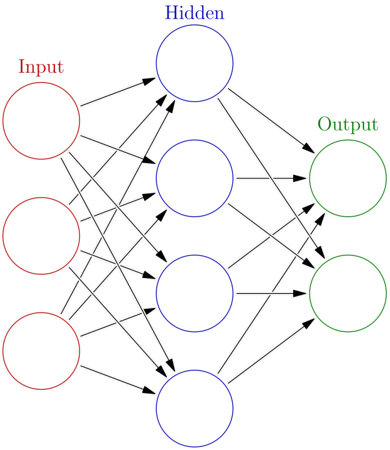
- Many useful networks have millions of weights
- A basic grid search for 1 million weights would take years
- For differentiable loss L, we can use the chain rule to efficiency calculate

$$\frac{\partial}{\partial w_i} L(\hat{f}_W(x) - f(x))$$

for all i (this is know as **backpropagation**)

$$-\left(\frac{\partial}{\partial w_1} L(\hat{f}_W(x)-f(x)), \ldots, \frac{\partial}{\partial w_n} L(\hat{f}_W(x)-f(x))\right)^\top$$

points in direction that decreases loss

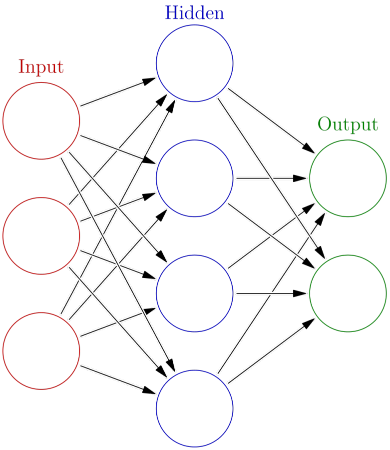


Source: Wikipedia

BASIC FEED FORWARD NEURAL NETWORKS – STOCHASTIC GRADIENT DESCENT

Algorithm 1 stochastic gradient descent neural network training

$\lambda \leftarrow$ learning rate
while you haven't lost patience **do**
    $x \leftarrow$ example from training set
    # Calculate the error
    $E \leftarrow \text{loss}(f(x), \hat{f}_W(x))$
    # Update the weights
    for weights w_i in network **do**
        $w_i \leftarrow w_i - \lambda \frac{\partial E}{\partial w_i}$
    end for
end while
Return: Trained neural network $\hat{f}_W$



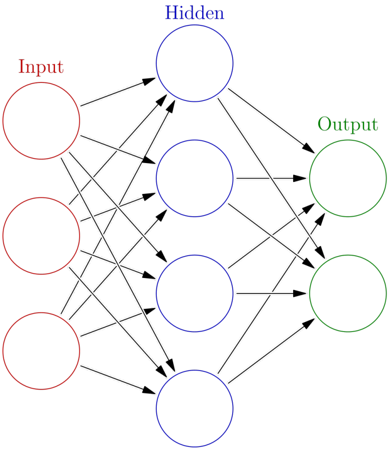
Source: Wikipedia

PUTTING IT ALL TOGETHER

- Neural networks can approximate arbitrary smoothish functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
- Given a value $x \in \mathbb{R}^n$, and a loss function L the gradients

$$\frac{\partial}{\partial w_i} L(\hat{f}_W(x), f(x))$$

  can be efficiently calculated for all weights w_i
- The weights W for network $\hat{f}_W$ can be learned on an arbitrarily large training set using Stochastic Gradient Descent
- All this can be done in parallel on the GPU



Source: Wikipedia

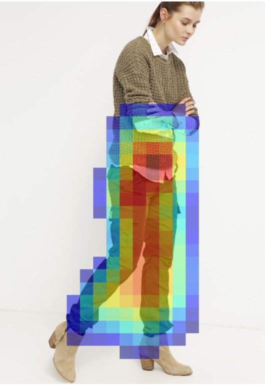
Semi-Supervised Localization

EXCITING SHOP THE LOOK APPLICATION: LOCALIZATION

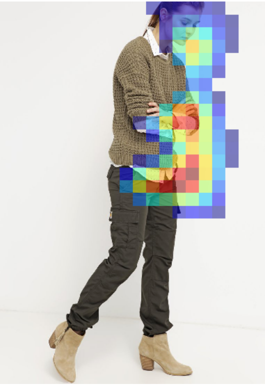
NO PRIOR INFORMATION ABOUT ARTICLE LOCATION WAS USED!

Pullover

Ankle Boots



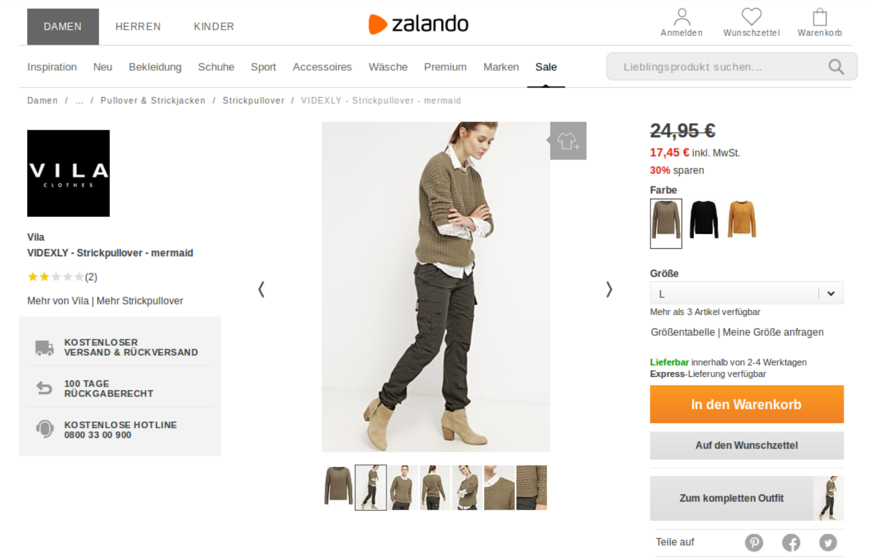
Trouser



Shirt

“SHOP THE LOOK FEATURE”

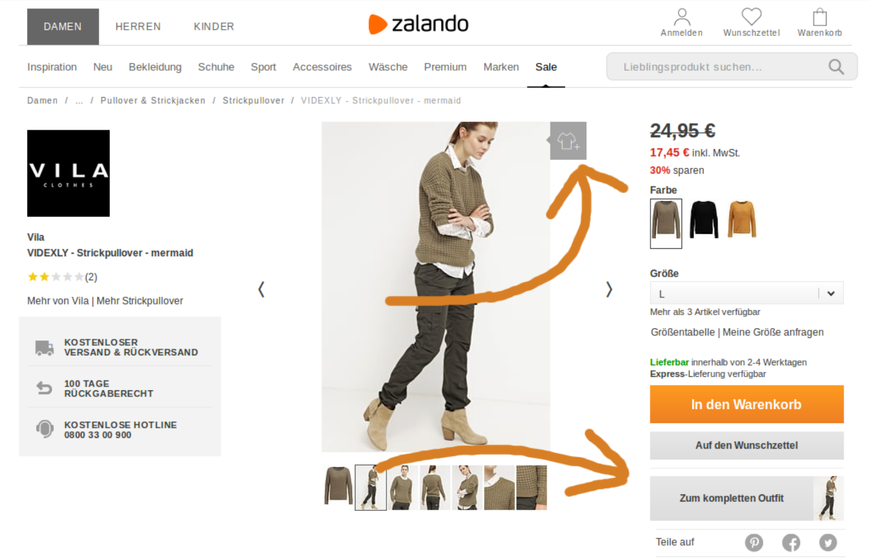
VIEW THE PRODUCT DETAIL PAGE



zalando research

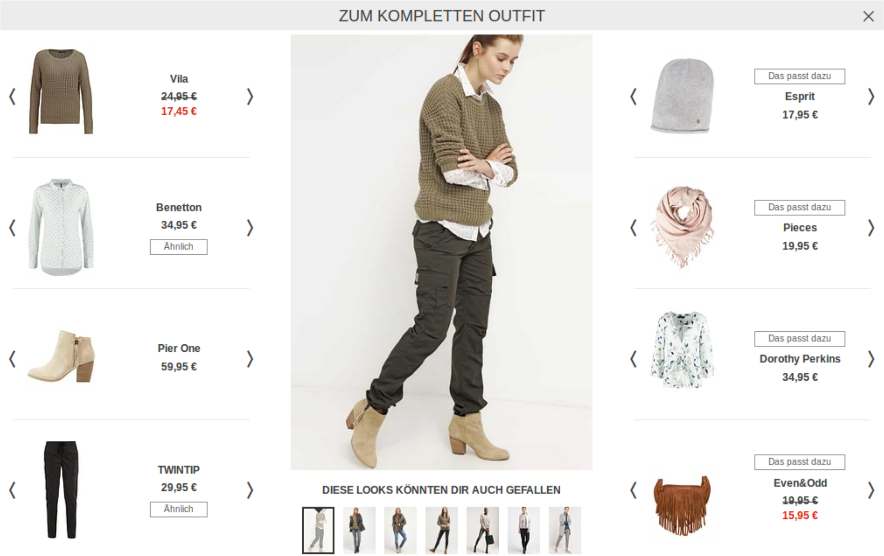
"SHOP THE LOOK FEATURE"

SAY YOU LIKE THE OTHER ITEMS



“SHOP THE LOOK FEATURE”

YOU CAN BUY THE OTHER ITEMS

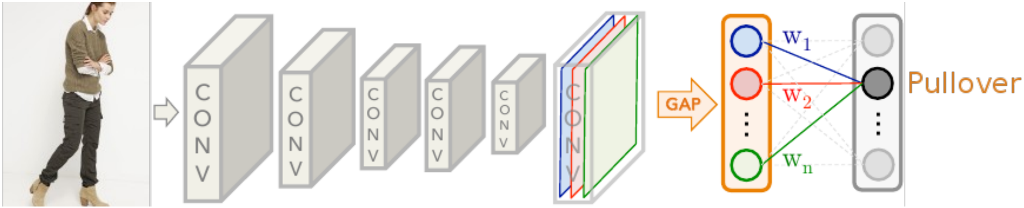


PRETTY COOL DATASET

Shop The Look Image

Individual Articles with Meta-data

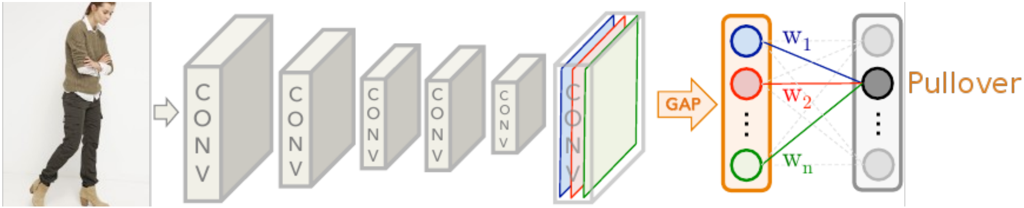
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

The method we used is based on *“Learning Deep Features for Discriminative Localization”* by Zhou, Khosla et. al, 2015. [4]
This image was also shamelessly pilfered from their paper.

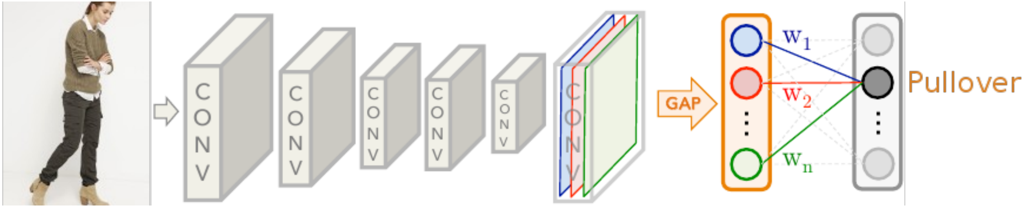
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

We start by training a classifier with a specific architecture (more on this later).

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



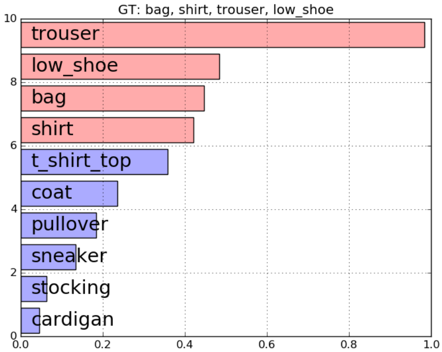
source: [4]

We start by training a classifier with a specific architecture (more on this later). For example, with this image inputted, output should be near

$$(0, 1, 0, 0 \ldots, 0, 1)^\top$$

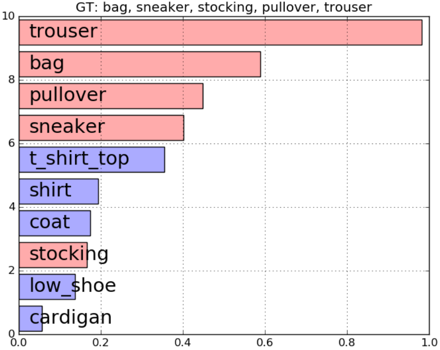
where the ones codify the articles that are present (pullover, ankle boots, trouser, shirt) and the zeros codify the articles missing (dress, sunglasses, ...).

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



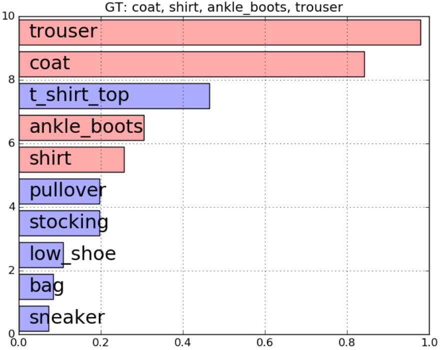
Classification Performance

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



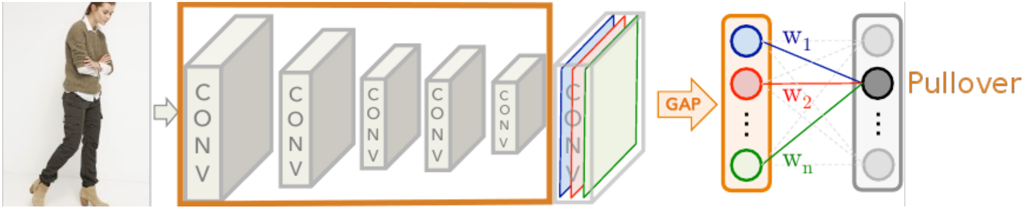
Classification Performance

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



Classification Performance

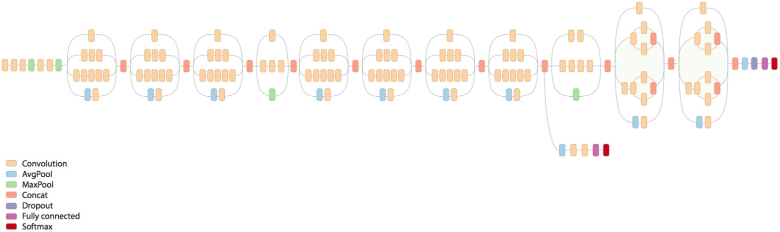
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

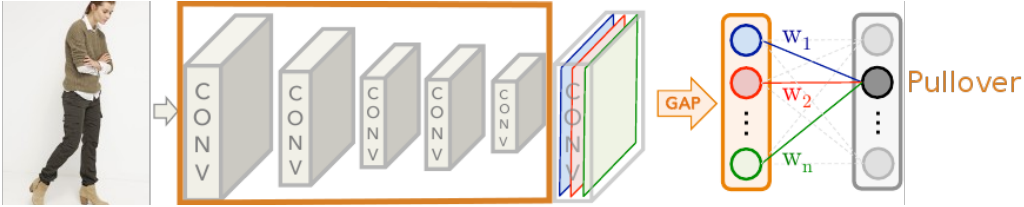
- Image is fed through most of pre-trained Google inception neural network

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: Deepmind

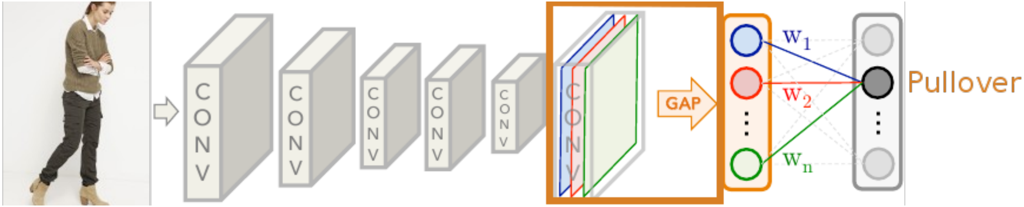
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

- Image fed through convolutional part of pre-trained inception network

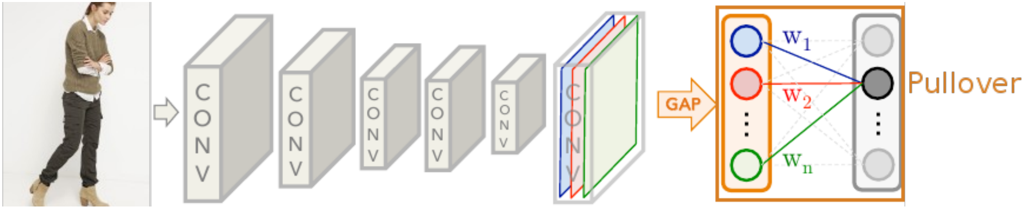
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

- Image fed through convolutional part of pre-trained inception network
- Results pooled with average pooling

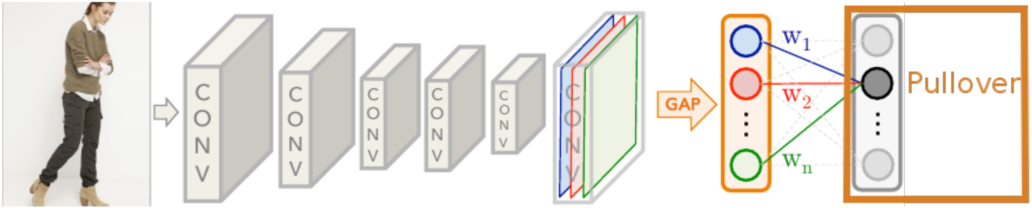
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

- Image fed through convolutional part of pre-trained inception network
- Results pooled with average pooling
- Pooled vector is transformed with a last fully connected matrix, giving us a classification

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION

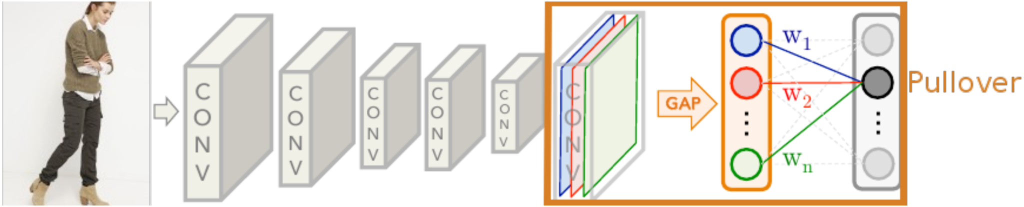


source: [4]

- Image fed through convolutional part of pre-trained inception network
- Results pooled with average pooling
- Pooled vector is transformed with a last fully connected matrix, giving us a classification
- Cross entropy loss back propagated

$$\sum_{i=1}^{n} \sum_{j=1}^{m} -\pi_{ij} \log(\sigma(c_{ij})) - (1 - \pi_{ij}) \log(1 - \sigma(c_{ij}))$$

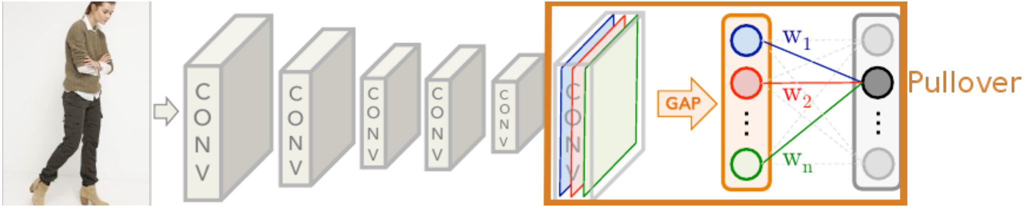
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

- Since the final bit is entirely linear, it's equivalent to an ensemble of classifiers
- Each classifier takes as input all channels of one spacial position of convolutional output
- Each individual classification depends on only a local patch of image
- Final classifier is average of individual classifiers

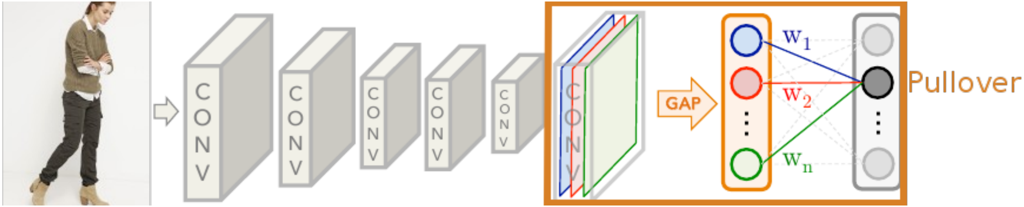
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

And that's how we manage to learn the location of fashion articles without any prior location information

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION



source: [4]

And that's how we manage to learn the location of fashion articles without any prior location information

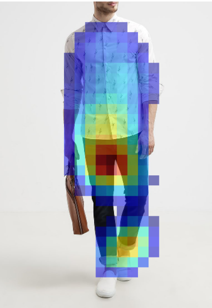
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION

A few more images from test set

Bag

Low Shoe

Shirt

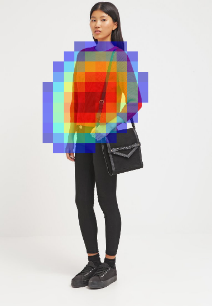


Trouser

DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION

A few more images from test set

Bag



Pullover

Sneaker

Trouser

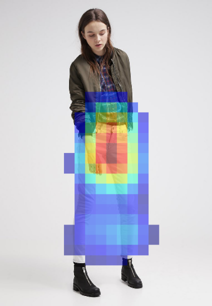
DEEP FEATURES FOR DISCRIMINATIVE LOCALIZATION

A few more images from test set


Ankle Boots


Coat


Shirt

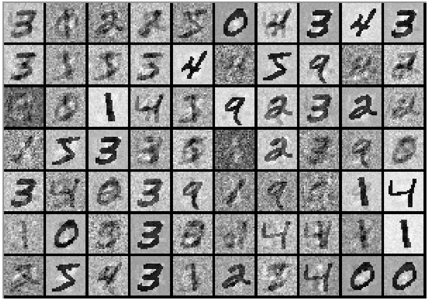

Trouser

Generative Adversarial Networks

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $G : \mathbb{R}^n \to \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_g such that if Z is some multivariate random noise

$$G : Z \to \{\text{cat pictures}\}$$



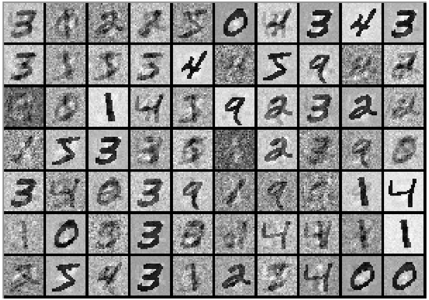
MNIST digits generated with restricted boltzmann machine. Source: https://deeplearning4j.org/rbm-mnist-tutorial.html

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $G : \mathbb{R}^n \to \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_g such that if Z is some multivariate random noise

$$G : Z \to \{\text{cat pictures}\}$$

- Two measures of quality:
  - Quality of generated images
  - Diversity of generated images



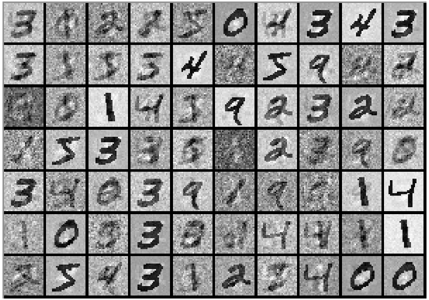
MNIST digits generated with restricted boltzmann machine. Source: https://deeplearning4j.org/rbm-mnist-tutorial.html

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $G : \mathbb{R}^n \to \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_g such that if Z is some multivariate random noise

$$G : Z \to \{\text{cat pictures}\}$$

- Two measures of quality:
  - Quality of generated images
  - Diversity of generated images
- Big challenge: How do you evaluate G with respect to these measures?



MNIST digits generated with restricted boltzmann machine. Source: https://deeplearning4j.org/rbm-mnist-tutorial.html

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $D : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_d such that

$$D : \{\text{pictures}\} \rightarrow \begin{cases} 1 & \text{It's a real cat picture} \\ 0 & \text{It's a generated cat picture} \end{cases}$$

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $D : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_d such that

$$D : \{\text{pictures}\} \rightarrow \begin{cases} 1 & \text{It's a real cat picture} \\ 0 & \text{It's a generated cat picture} \end{cases}$$

- If G creates poor quality images, D will detect it
- If G fails to create diverse images, D will learn which images are generated

GENERATIVE ADVERSARIAL NETWORK MOTIVATION

- Neural network $D : \mathbb{R}^n \to \mathbb{R}^m$ is a universal function approximation.
- There must exist weights W_d such that

$$D : \{\text{pictures}\} \to \begin{cases} 1 & \text{It's a real cat picture} \\ 0 & \text{It's a generated cat picture} \end{cases}$$

- If G creates poor quality images, D will detect it
- If G fails to create diverse images, D will learn which images are generated
- $D(G(z))$ is a neural network, so we can use gradients from D to update G

CLASSIC GENERATIVE ADVERSARIAL NETWORK FORMULATION

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{x \sim p_z(z)}[\log(1 - D(G(z)))]$$

CLASSIC GENERATIVE ADVERSARIAL NETWORK FORMULATION

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{x \sim p_z(z)}[\log(1 - D(G(z)))]$$

Algorithm 3 GAN training algorithm from [2]

for number of training iterations **do**
    for k steps **do**
        Sample noise vector $z \sim p_g(z)$ and real data $x \sim p_{\text{data}}(x)$
        Update discriminator by using its gradient:

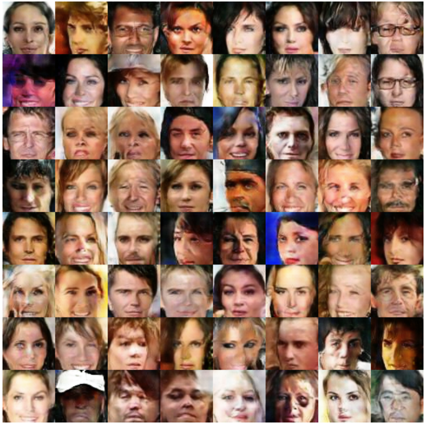
$$\nabla_{W_d} \log D(x) + \log(1 - D(G(z)))$$

    end for
    Sample noise vector $z \sim p_g(z)$
    Update generator by using its gradient:

$$\nabla_{W_g} \log(1 - D(G(z)))$$

end for

GANS FOR CREEPY FACES

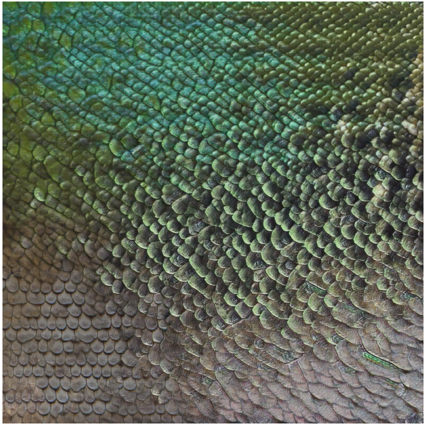


Faces generated by DCGAN trained on CelebA dataset [3]

GANS FOR FASHION

Interpolation between random items in fashion DNA space

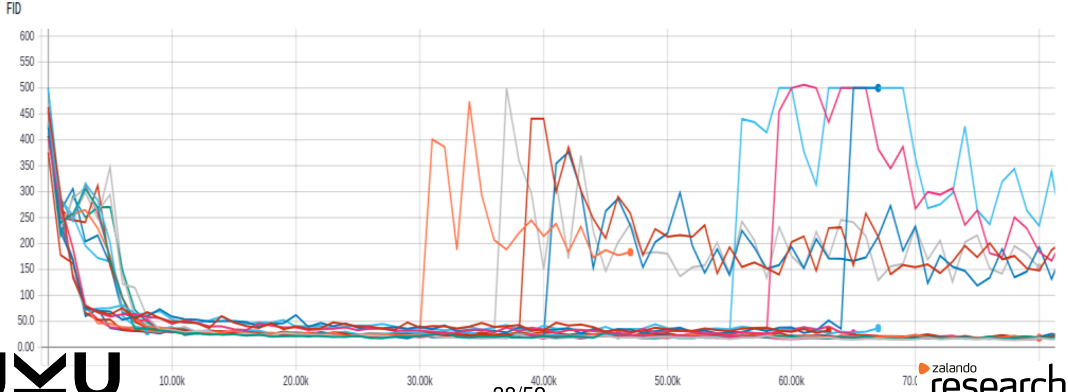
GANS FOR TEXTURE GENERATION



Generated texture where each corner is a specific snake skin texture and interior is a linear combination of corner textures, see [1]

GAN OPEN QUESTIONS / PROBLEMS

- Convergence is the exception, not the rule
- Many successful GANs are highly tuned
- Finding exact right parameters seems to be trial and error
- Sudden divergence

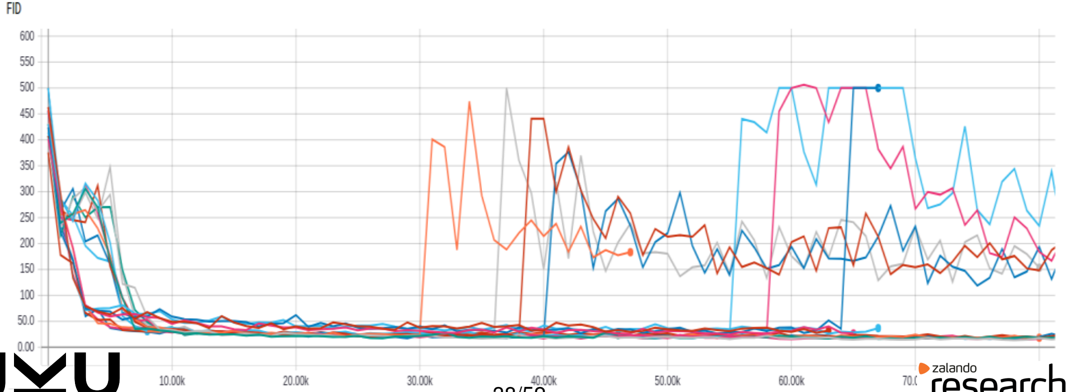


GAN OPEN QUESTIONS / PROBLEMS

- Convergence is the exception, not the rule
- Many successful GANs are highly tuned
- Finding exact right parameters seems to be trial and error
- Sudden divergence

GAN convergence sensitive to:

- Learning rates
- Activation functions
- Generator / discriminator architectures
- Alignment of the stars ;)



The Most Important Part of the Talk

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

In his book, Shawn Otto outlines how the regulatory environment adapts to new technologies. Think things like pesticides and insecticides, fossil fuel consumption, opioid pain killers.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

A new process or tool (for example, a chemical or, today, a nanotechnology or genetic technology) is discovered that vastly expands utility, power, convenience, or efficiency.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Industrial applications are quickly developed and commercialized, often increasing productivity and lowering costs. But the science of biocomplexity and ecology—of how the process or tool will affect and be affected by its broader context, from the human body to the environment—lags behind.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Industries grow up around the new application. Major capital investments are made and its use intensifies.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

A tipping point is reached at which the application has noticeable negative effects on health or the environment. Fueled by growing public outcry, scientists study the degree of the systemic effect to determine what to do.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Regulations proposed to minimize the negative effects, but vested economic interests sense a potentially lethal blow to their production systems and fight the proposed changes by denying the environmental effects, maligning and impeaching witnesses, questioning the science, attacking or impugning the scientists, and/or arguing that other factors are causing the mounting disaster. A battle ensues between the adherents of old science and those of new science.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Evedince continues to accumulate from the emerging science until the causation becomes irrefutable, often through dramatic deaths or disasters (or in the case of climate disruption, extreme weather events) that draw increased public scrutiny and outrage, finally tipping the politics in the direction of reform.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Regulations are passed or laws are changed to stop or modify use and to mitigate the effects. The industrial approach grudgingly shifts to take into account the relationships between the application and its environmental and/or physiological context. Or this does not occur, in which case the process returns to stage 3.

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

Where do you see artificial intelligence here?

How can artificial intelligence become part of society in a beneficial way?

THE SEVEN STAGES OF TECHNOLOGICAL ADAPTATION

This text is from *The War on Science* by Shawn Otto

1. **Discovery**
2. **Application**
3. **Development**
4. **Boomerang**
5. **Battle**
6. **Crisis**
7. **Adaptation**

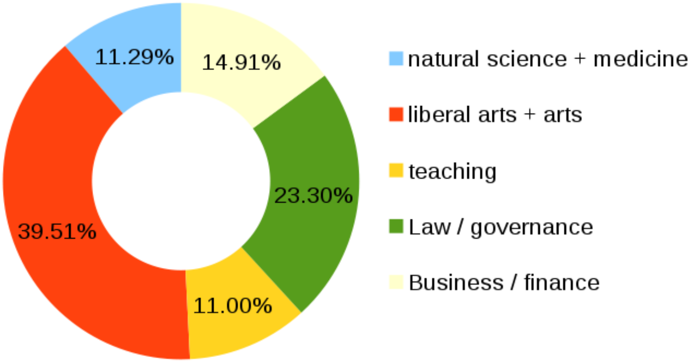
Where do you see artificial intelligence here?

How can artificial intelligence become part of society in a beneficial way?

The answers are both political and scientific

How can you bring science into the political discourse?

SCIENTISTS IN THE BUNDESTAG



Percent of subjects studied by MPs of the 18th German Bundestag

*These numbers are hard to calculate exactly, see backup for full methodology

Thanks for listening

REFERENCES

Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf.
Learning texture manifolds with the periodic spatial gan.
arXiv preprint arXiv:1705.06566, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Alec Radford, Luke Metz, and Soumith Chintala.
Unsupervised representation learning with deep convolutional generative adversarial networks.
arXiv preprint arXiv:1511.06434, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba.
Learning deep features for discriminative localization.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

abgeordnete

			natural science + medicine		liberal arts + arts		teaching		Law / governance		Business / finance		engineering	total
Anglistik	11		0	1	11		0		0		0		0	1
Architektur	3		0		0		0		0		0	1	3	1
Biologie	13	1	13		0		0		0		0		0	1
Chemie	6	1	6		0		0		0		0		0	1
Geografie / Geologie	7	1	7		0		0		0		0		0	1
Germanistik	19		0	1	19		0		0		0		0	1
Geschichte	44		0	1	44		0		0		0		0	1
Gesellschaftswissenschaften	2		0	1	2		0		0		0		0	1
Ernährungs- und Haushaltswissenschaften	1	1	1		0		0		0		0		0	1
Informatik	5	1	5		0		0		0		0		0	1
Ingenieurwesen	22		0		0		0		0		0	1	22	1
Islamwissenschaften	1		0	1	1		0		0		0		0	1
Journalistik, Publizistik	3		0	1	3		0		0		0		0	1
Kulturwissenschaften	4		0	1	4		0		0		0		0	1
Kunstgeschichte	5		0	1	5		0		0		0		0	1
Landwirtschaft, Forstwirtschaft	8		0		0		0		0	1	8		0	1
Lehramt / Dipl.-Lehrer(in)	38		0		0	1	38		0		0		0	1
Literaturwissenschaften	5		0	1	5		0		0		0		0	1
Mathematik	10	1	10		0		0		0		0		0	1
Medien- und Kommunikationswissenschaften	12		0	1	12		0		0		0		0	1
Medizin	8	1	8		0		0		0		0		0	1
Musikwissenschaften	5		0	1	5		0		0		0		0	1
Orientalistik	1		0	1	1		0		0		0		0	1
Pädagogik	31		0		0	1	31		0		0		0	1
Pharmazie	1	1	1		0		0		0		0		0	1
Philologie, Philosophie	15		0	1	15		0		0		0		0	1
Physik	8	1	8		0		0		0		0		0	1
Politikwissenschaften	79		0	1	79		0		0		0		0	1
Psychologie	7	1	7		0		0		0		0		0	1
Rechts- und Staatswissenschaften	149		0		0		0	1	149		0		0	1
Romanistik	6		0	1	6		0		0		0		0	1
Sozialarbeit	7		0		0	1	7		0		0		0	1
Sozialwissenschaften	14		0	1	14		0		0		0		0	1
Soziologie	31		0	1	31		0		0		0		0	1
Sportwissenschaften	7	1	7		0		0		0		0		0	1
Sprachwissenschaften	1		0	1	1		0		0		0		0	1
Theologie	13		0	1	13		0		0		0		0	1
Umweltwissenschaften	1	1	1		0		0		0		0		0	1
Verwaltungswissenschaften	12		0		0		0	1	12		0		0	1
Veterinärmedizin	4	1	4		0		0		0		0		0	1
Volkskunde	2		0	1	2		0		0		0		0	1
Volkswirtschaft	33		0		0		0		0	1	33		0	1
Wirtschafts- und Sozialwissenschaften, Betriebswirtschaft	62		0		0		0		0	1	62		0	1
http://www.bundestag.de/abgeordnete18/mdb_zahlen	716		78		273		76		161		103		25	716

natural scienc	0.108938547
liberal arts + a	0.381284916
teaching	0.106145251
Law / governa	0.224860335
Business / fina	0.143854749