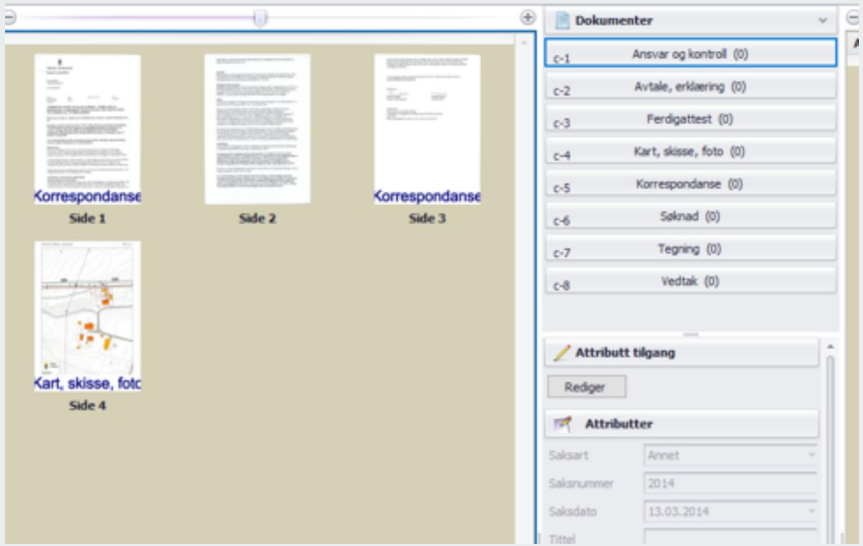
Bacheloroppgave 2019

Dataingeniør NTNU

Oppgaven



- Gruppere sammen sammenhengende sider
- Forbedre klassifiseringsmodell

Hvorfor valgte vi denne oppgaven?

Metode

- Brainstorming

- Liste over idéer

Prøv og feil

- Term Frequency-Inverse Document Frequency
- Teksturanalyse
- Sidetall
- Dato
- Sidestørrelse

Prøv og feil

- Term Frequency-Inverse Document Frequency
- Teksturanalyse
- Sidetall
- Dato
- Sidestørrelse
- Doc2Vec
- Gjennomsnittsfarge
- N-gram
- Jaccard similarity
- Kombinasjoner

Resultat

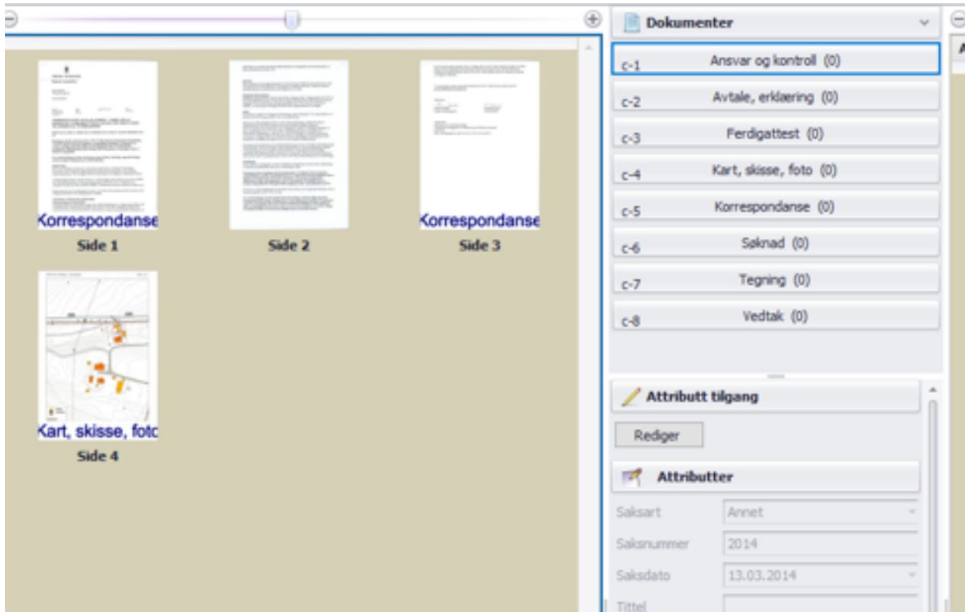
	Klepp L.	
Modell	TP	FP
N-gram	16 567	1 105
Jaccard	10 936	882
Sidetall	631	1
Sidestørrelse	25 017	1 011

Resultat

	Klepp L.	
Modell	TP	FP
N-gram	16 567	1 105
Jaccard	10 936	882
Sidetall	631	1
Sidestørrelse	25 017	1 011
Kombinert	36 804	2 484
PPV	93.7%	

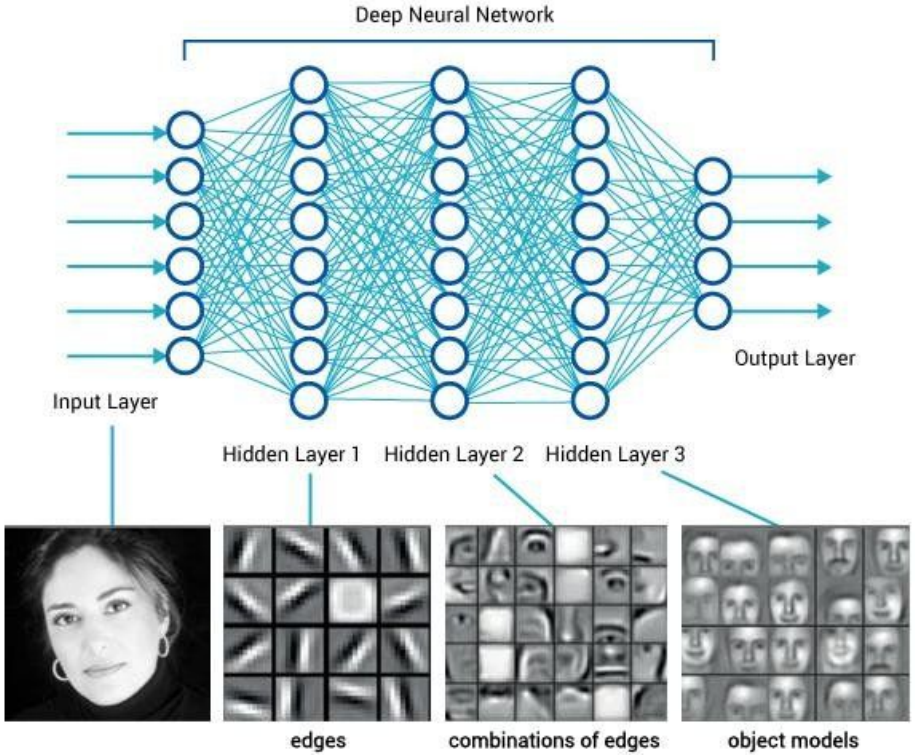
Resultat

	Klepp	Klepp L.	Hamar	Tromsø	Average
Classifier	48.3%	53.2%	**56.9%**	**73.4%**	**58.0%**
NC	48.3%	53.1%	56.4%	72.2%	57.5%
NCR	**48.4%**	53.2%	56.5%	72.3%	57.6%
NC+	**48.4%**	53.2%	55.8%	72.2%	57.4%
NCR+	**48.4%**	**53.3%**	56.0%	72.3%	57.5%

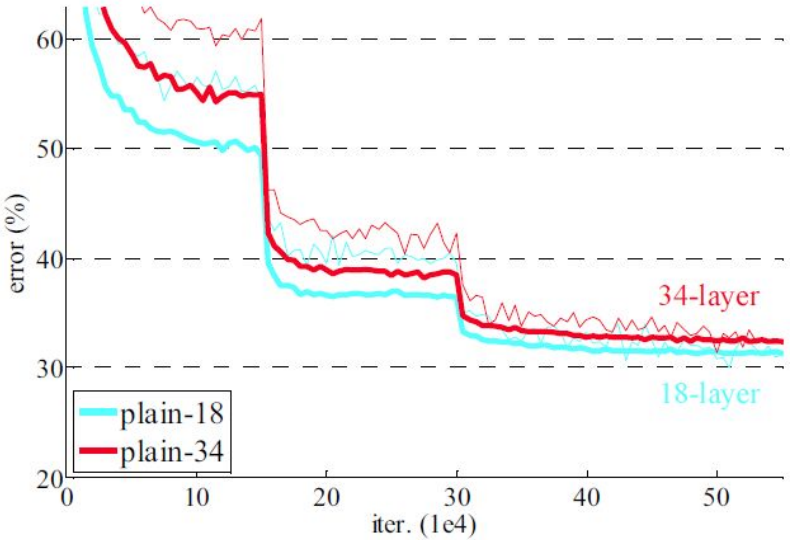

Korrespondanse
Side 1
Side 2
Korrespondanse
Side 3
Kart, skisse, foto
Side 4
Dokumenter
c-1 Ansvar og kontroll (0)
c-2 Avtale, erklæring (0)
c-3 Ferdigattest (0)
c-4 Kart, skisse, foto (0)
c-5 Korrespondanse (0)
c-6 Søknad (0)
c-7 Tegning (0)
c-8 Vedtak (0)
Attributt tilgang
Rediger
Attributter
Saksart
Annet
Saksnummer
2014
Saksdato
13.03.2014
Tittel

Bildemodell

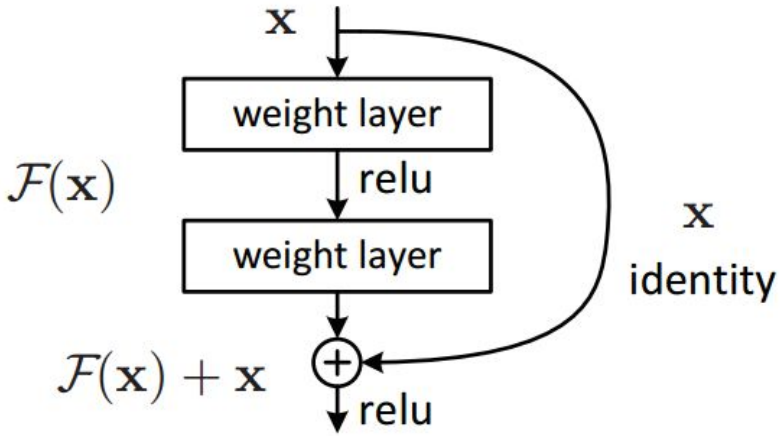
- Neural network 101
- Matriseoperasjoner
- Justere vekter slik at input gir rett output
- Håper dette generaliserer



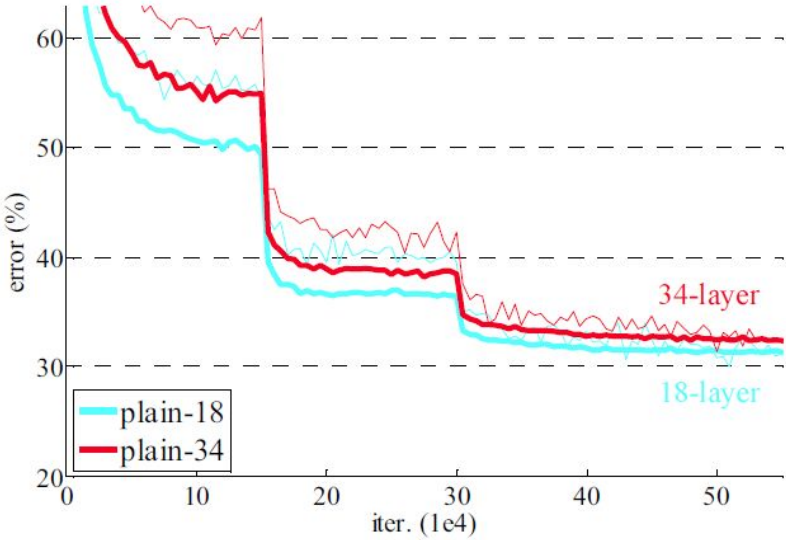
Problem

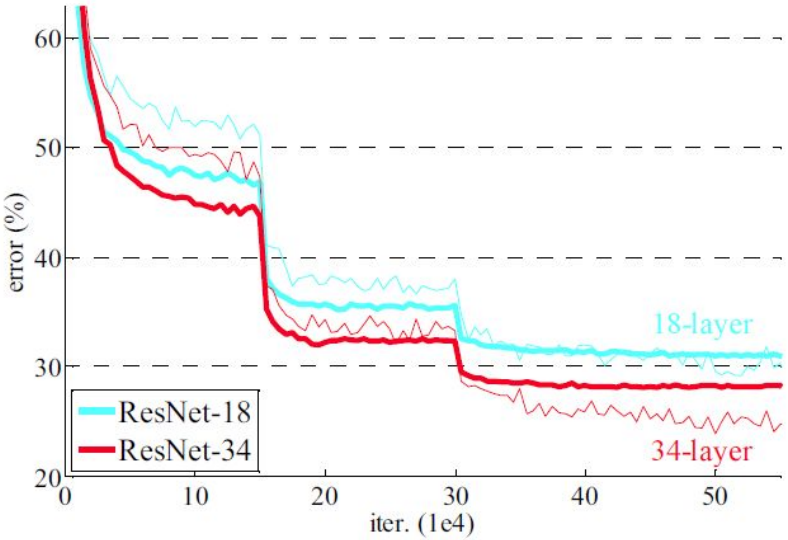

60
50
40
30
20
error (%)
0
10
20
30
40
50
iter. (1e4)
34-layer
18-layer
plain-18
plain-34

Løsning: Residual block



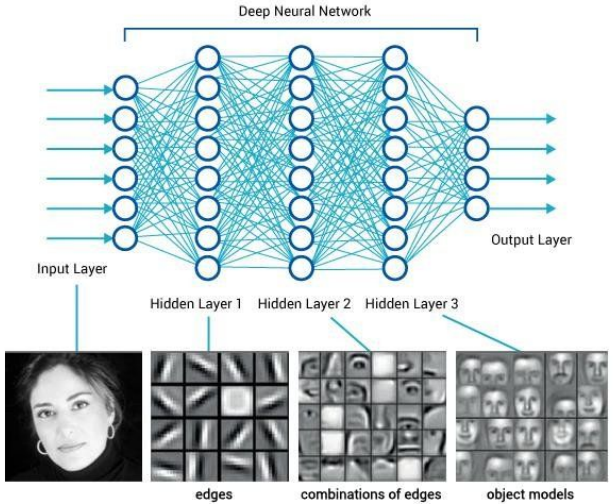
Effekt


60
50
40
30
20
error (%)
0
10
20
30
40
50
iter. (1e4)
34-layer
18-layer
plain-18
plain-34


60
50
40
30
20
error (%)
0
10
20
30
40
50
iter. (1e4)
18-layer
34-layer
ResNet-18
ResNet-34

Vår løsning

- ResNet50 (50 layers med residual blocks)
- Tilpasset output til hvert datasett
- Til sammen ca. 23,5 millioner vekter
- Trener på et offentlig arkiv
  - Transfer learning
  - Kraftigere hardware



Resultat

	Klepp	Klepp L.	Hamar	Tromsø	Average
ResNet	40.8%	63.8%	37.6%	68.9%	52.8%

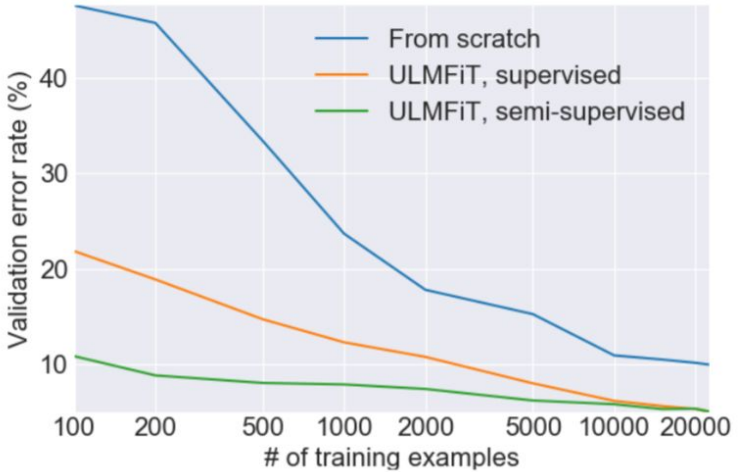
Hva holder oss tilbake?

- Begrenset maskinkraft til trening
- Lite treningsdata
- Klassefordeling
  - Klepp L.: 54 203 Annet, 8 Korrespondanser
  - Noen klasser er vanskelige å skille (“Annet”, “Kart, tegning, foto”, “Tegning, foto”)
  - Standardiserte klasser → generell modell for alle arkiv

90.6% på et annet offentlig arkivdatasett (RVL-CDIP)

ULMFiT

- Universal Language Model Fine-tuning for Text Classification
- Transfer learning for tekst

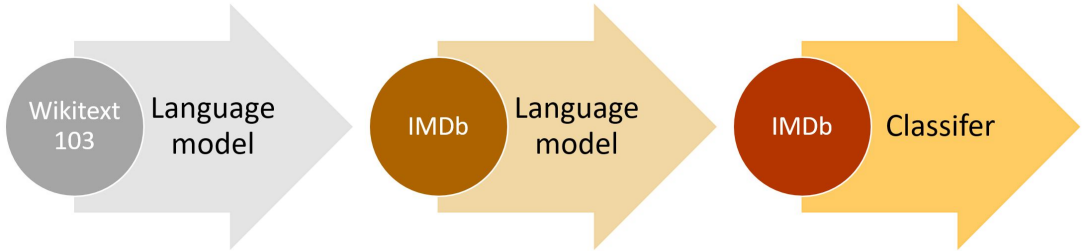


Språkmodellen

- OCR fra dokumenter
- Tokenize
- Vokabular
- LSTM modell
- Gjetter neste ord

Transfer learning

- Samme prinsipp som bilde
- Wikipedia
- Finjustering
- Vektene brukes i klassifiseringsmodell



Resultater

	Klepp	Klepp L.	Hamar	Tromsø	Average
Existing classifier	10.8%	62.4%	69.7%	71.1%	53.4%
ULMFiT	81.7%	91.1%	77.8%	88.2%	84.7%
ResNet	40.8%	63.8%	37.6%	68.9%	52.8%

Videre arbeid

- Transfer learning mellom arkiv
  - Eventuelt standardisere klasser
- Fintune modeller mer (krever maskinkraft)
- Kombinere tekst- og bildemodellene
- Active learning
  - Modellen spør når den er usikker
  - Trener inn slik at den blir mer og mer nøyaktig kontinuerlig

Spørsmål?