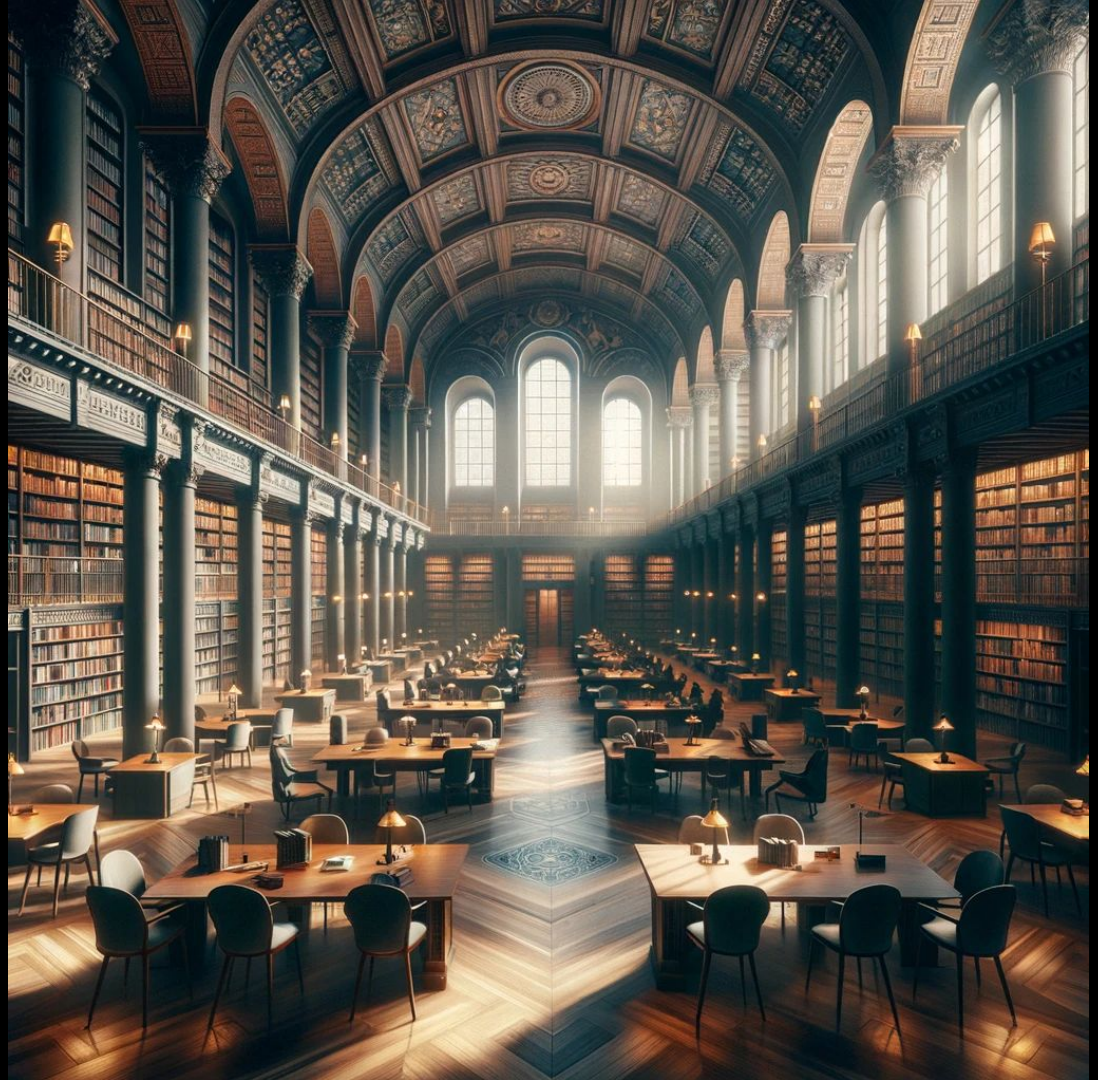
NB-Whisper

Per Egil Kummervold
per.kummervold@nb.no
AI-lab
National Library of Norway



LTG Master - 22.04.24



illustrasjon: ChatGPT





Arkivfoto Nasjonalbibliotektet

ZDNET

Norway's petabyte plan: Store everything ever published in a 1,000-year archive

From ancient manuscripts to movies, the National Library of Norway wants to put it all online for the public.



Started 2008

Status August 2022

88%	724.000	books
75%	5.180.000	newspapers
4%	2.000.000	magazines
50%	3.391.000	photos
58%	1.083.000	radio hours

How to solve “speech to text”?

- Lets convert phonems to graphemes!

How to solve “speech to text”?

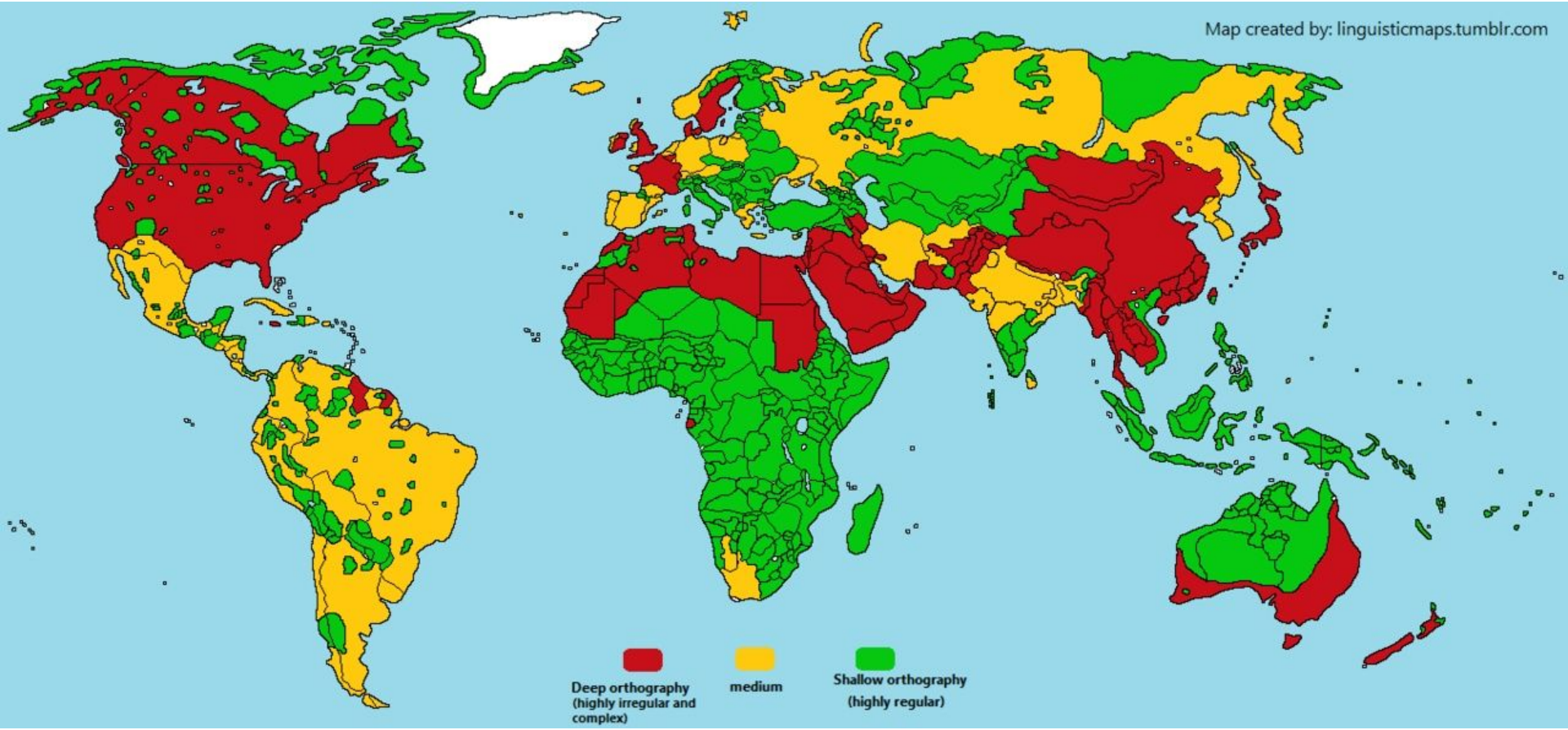
- Lets convert phonems to graphemes!

- We have been trying this for 70 years

How to solve “speech to text”?

- Lets convert phonems to graphemes!
- We have been trying this for 70 years
- It does not work!

WHY?


Map created by: linguisticmaps.tumblr.com
Deep orthography (highly irregular and complex)
medium
Shallow orthography (highly regular)

Norwegian - Orthographic Depth

Tekst	**IPA - lyder**
jeg	/jæɪ/
er	/æːr/
veldig	/ˈʊɛldɪ/
sliten	/ˈʂliːtən/

Norwegian - Orthographic Depth

Tekst	Oslo	Nordmøre
jeg	/jæɪ/	/ɛ/
er	/æːr/	/ɛ/
veldig	/ˈʊɛldɪ/	
sliten	/ˈʂliːtən/	

Norwegian - Orthographic Depth

Tekst	Oslo	Nordmøre
jeg	/jæɪ/	/ɛ/
er	/æːr/	/ɛ/
veldig	/ˈʊɛldɪ/	/ˈçøːɽə/
sliten	/ˈʂliːtən/	/kɽɑːr/

Second Reason - No one really wants it!

“Men den debatt vi har hatt i samfunnet i de siste fem årene, faktisk, og mer enn det, ja, fem årene, for det var fra niogåtti, fire år, har jo blant annet dreiet seg om at noe av det folk, det har vært mulig å oppvise engasjement på, har vært hva man mener om EF, spesielt hvis man er mot, altså engasjement mot EF har vært noe av det engasjement som har vært mobilisert faktisk, og der har ikke kvinnene vært noe mindre engasjert enn menn i det.”

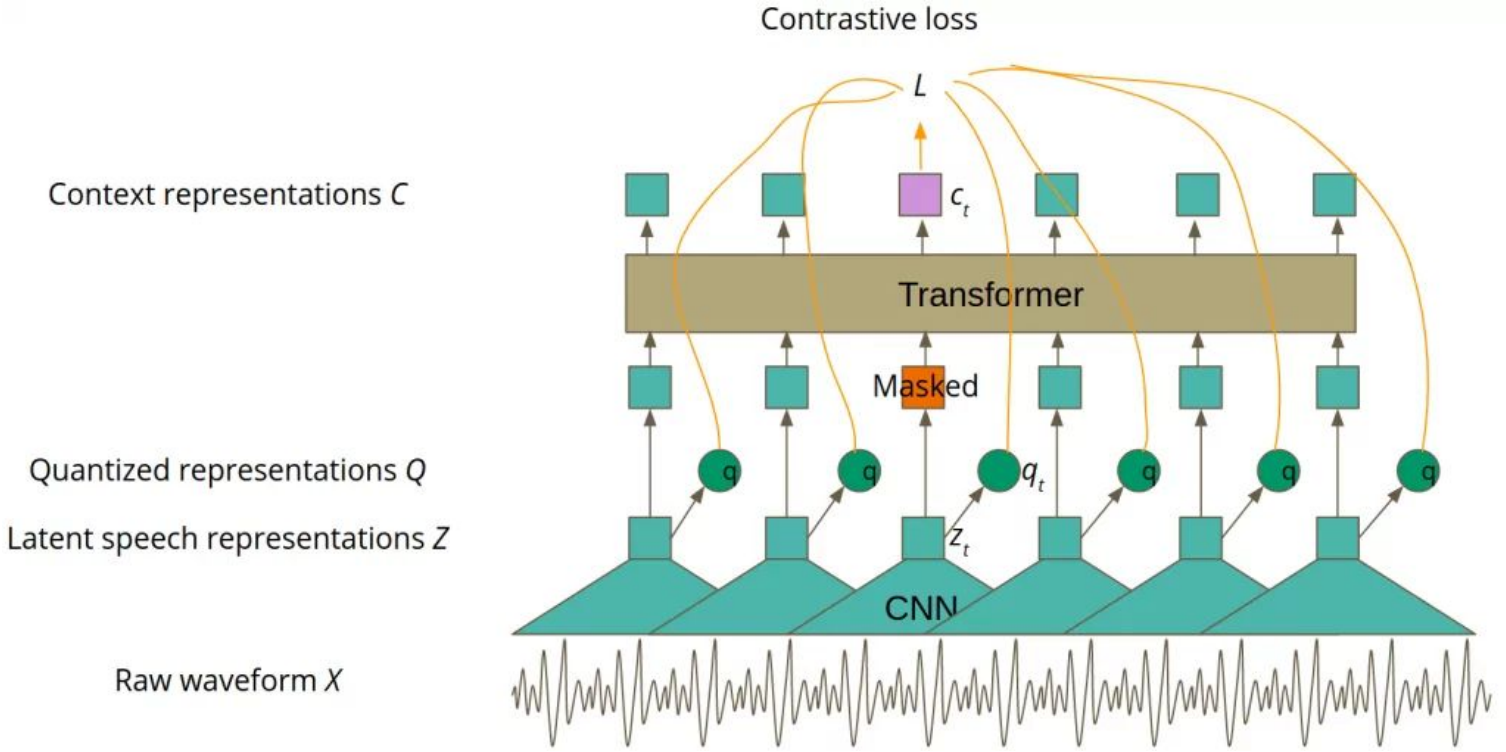
Gro Harlem Brundtland
Kapital 13/1993 - Niels Christian Geelmuyden

Our Starting Point

- Verbatim Transcription is not possible
- Verbatim Transcription is not really wanted
- Transcription is really a translation task
- There is not a single correct answer
- Transcription style depends on context

Alternatives - Wav2Vec2

- Encoder only-architecture
- Main idea is to create rich, contextual representations of audio
- Using convolutional networks to convert these audio segments to latent representations
- Pretrained a bit like BERT
- Pretrained on 960 hours of unlabeled data
- Can be finetuned on very little data (a few hours)
- Unlabeled data is cheaper than labeled data
- Typically post-processed to get capitalisation and punctuation



Source: Neurosys.com

Whisper

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

Abstract

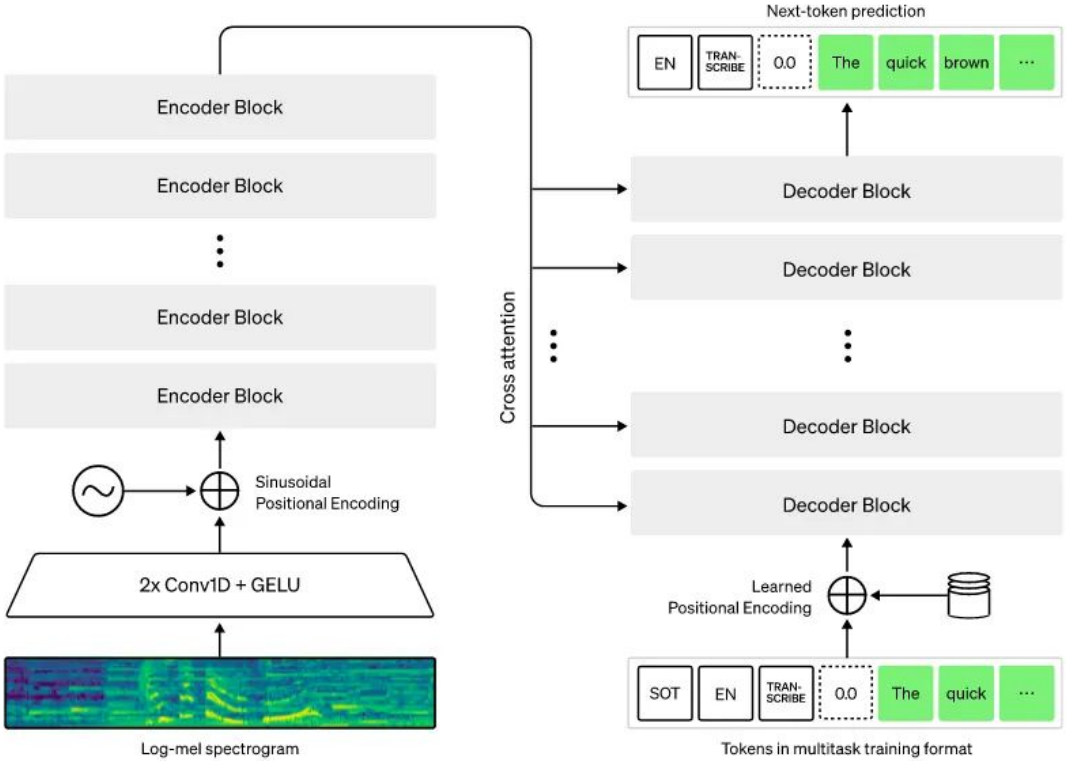
We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy

OpenAI, 2022

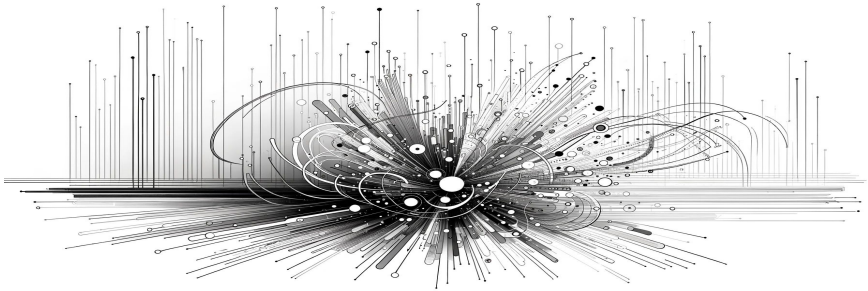
What is New?

- Proves everyone wrong about how these models can be trained
- Trained on “weak alignment”
- What is “weak alignment”
- Dropping the self-supervised pretraining
- Trained directly on 600k of subtitles
- Large multilingual models are better than large English models
- Follow the same training regime to finetune



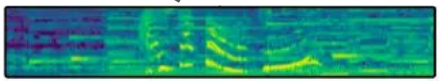
illustrasjon: OpenAI artikkel

multidimensional context vectors



encoder

decoder



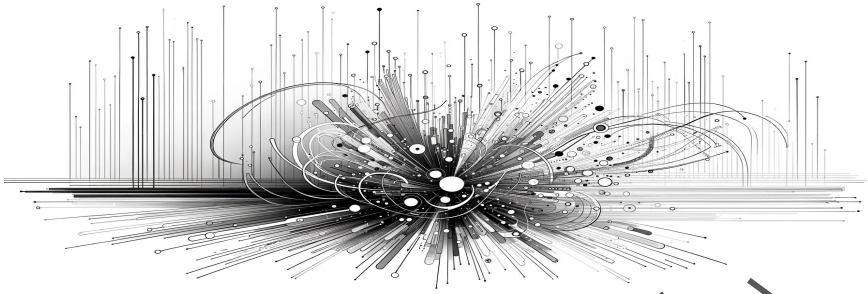
Log-mel spectrogram



Tokens in multitask training format

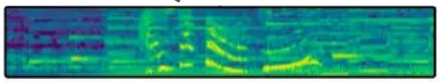
illustrasjon: OpenAI artikkel + ChatGPT

multidimensional context vectors



encoder

decoder



Log-mel spectrogram



Tokens in multitask training format





OpenAI Whisper Architecture + ChatGPT

Training Data - NB-Whisper

Datasett	Fase 1 (timer)
NRK - Undertekster	16 518
NRK - Uten undertekst	715
Lydbøker	2 461
NST-datasettet	260
Stortingets talekorpus	2 230
Totalt	**22 184**

- Up to 30 second audio samples
- More than 8 million training samples

Discuss

- What kind of error can be in such datasets?
- What could be the consequences of errors?
- How can it be cleaned?

Methods that does not work

- You can not solve all problems by adding more data
  - We have enough data
- You can not listen to 20k hours of audio
- Any form of manual editing just fails
- You can not solely train your model on the models output

Methods that works

- Transcription from multiple models
- Targeted editing
  - «Undertekster av …»/«Subtitles by ..»
  - «Next program in this series will be aired..»
  - »This program might contain scenes that are…»
- First and last word needs to be at least partially correct
- Detecting harmful insertions. N-grams. Typically 3-grams that does not exist in any prediction are deleted.
- Detecting harmful omissions. Also N-grams. Typically 3-grams in exists in all (or most) of the prediction, leads to the entire audio snippet is deleted.
- NER analysis to detect added names
-

Types of Data Errors

Challenges	Description	Positive Effects	Negative Effects
Style Variation	Differences in how data is presented, such as variations in style of transcript, for instance how verbatim the transcription is.	Can make the model more robust.	Can cause the model to change between these styles randomly. Negative effect on WER-scores.
Orthographic Variation	Differences in format, such as number representation, abbreviations, standardising orthographic variations etc.	Can make the model more robust.	Leads to non desirable variation in the output. Can lead to inconsistent model training.
Misalignment	Subtitles or transcribed text not accurately syncing with spoken words.	None	Major reason for hallucination. Clear negative effect.
Untranscribed Audio	Relevant or seemingly irrelevant spoken words not captured in transcriptions.	Omitting irrelevant interruptions can streamline the output. In most transcription styles, omitting verbal repetitions makes the text easier to read.	Might cause the model to neglect important content.
Extra Transcribed Text	Text added to the transcription that does not appear in the audio, including annotations about speakers or actions.	None	Major reason for hallucination. Clear negative effect. Especially problematic that names often are added to transcripts, since this leads to the model hallucinating names.

How are they trained?

- Basically two columns of data. Audio and text
  - Augmented for translation and timestamps
- Systematic data errors very problematic
  - Example from Whisper paper
  - Mr President
  - Proper names
- Data cleaning 80% of the job
-

Lessons Learned

- Whisper Large is tricky to train
  - Possibly over-parameterised
  - Needs large batch sizes to regularise
- Tiny, Base, Small and Medium trains easily and can easily be finetuned for a domain
- We needed to add dropout (especially for Large)
  - BPE dropout
  - Activation dropout
- Improving data quality have huge effect on result
- Evaluation of non-verbatim models are tricky
  - Very easy to get fooled by the metrics
- Style is all you need

Modellstørrelser

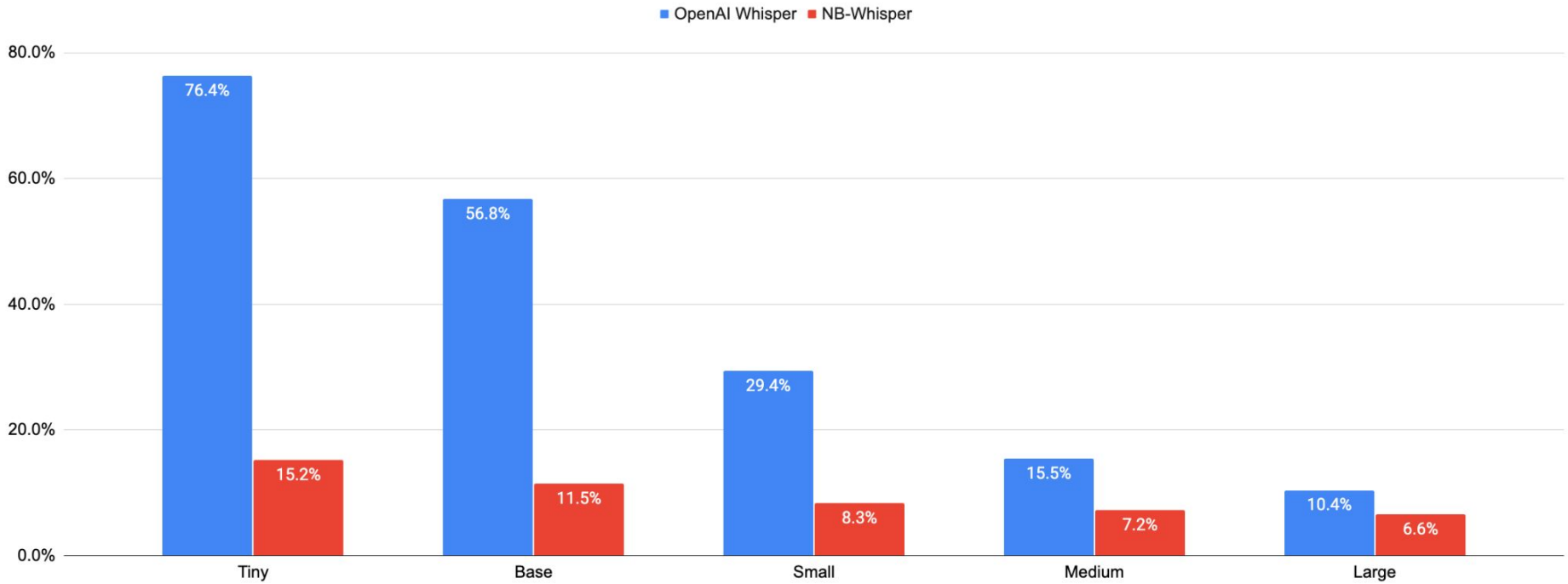




Størrelse	Parametre	VRAM	Hastighet
Tiny	39 M	~1 GB	~32x
Base	74 M	~1 GB	~16x
Small	244 M	~2 GB	~6x
Medium	769 M	~5 GB	~2x
Large	1550 M	~10 GB	1x



Fleurs - Bokmål



Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges - Preview - https://arxiv.org/abs/2402.01917

External Evaluation - WER

Bokmål	WER %
NB-Whisper Verbatim	11,89
NB-Whisper	18,71
Google USM	20,84
NB-Wav2Vec2	22,32
Microsoft Azure	24,63
Open AI Whisper	30,17
Google Cloud long	30,40
NB-Whisper Semantic	31,53

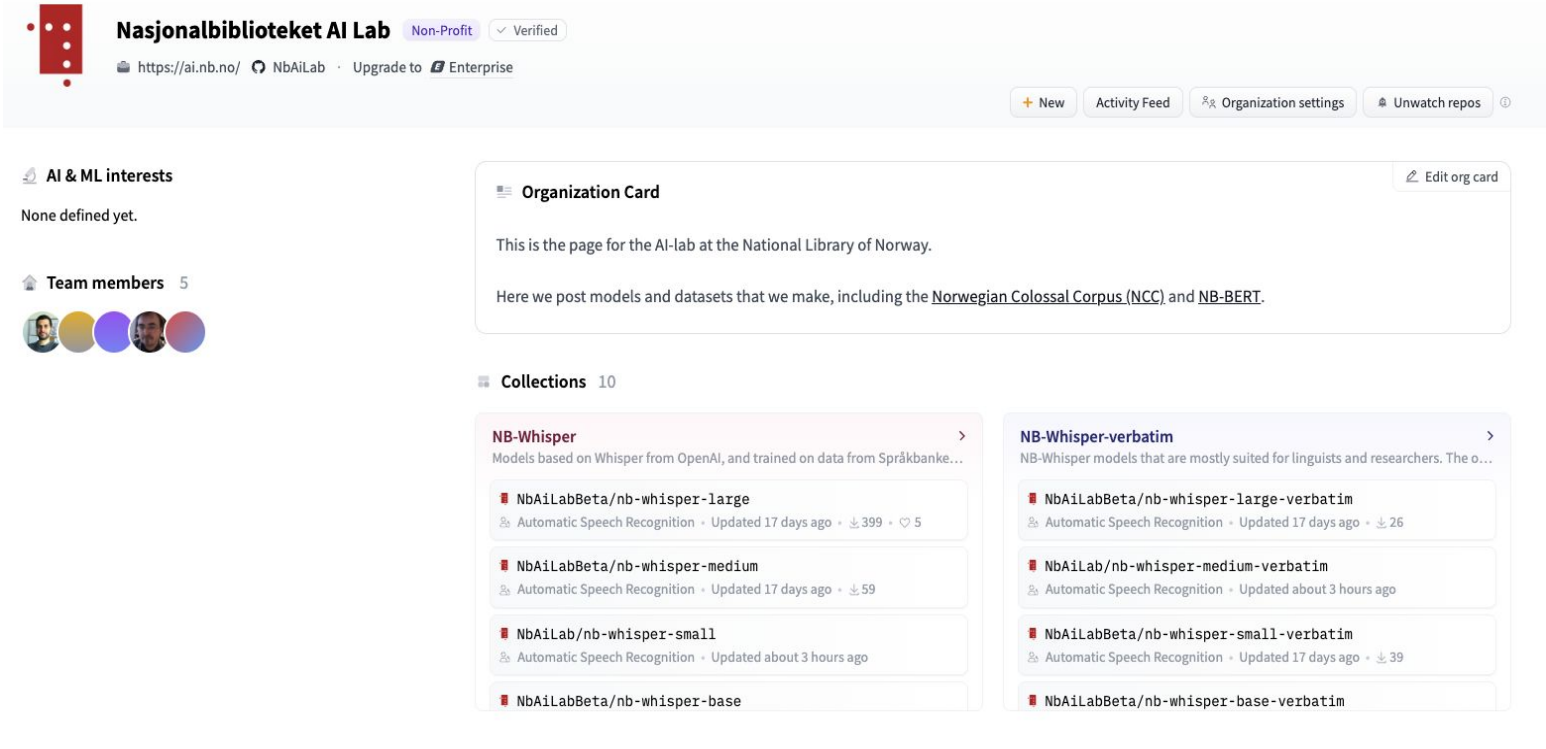
External Evaluation - SemDist

Bokmål	SemDist
NB-Whisper	0,09
NB-Whisper Semantic	0,11
NB-Whisper Verbatim	0,12
Microsoft Azure	0,17
OpenAI Whisper	0,20
NB-Wav2Vec2	0,21
Google USM	0,26
Google Cloud long	0,38

Dialekt bokmål

Bokmål WER %	Øst	Sørvest	Vest	Trøndersk	Nord
NB-Whisper Verbatim	**13,23**	**11,78**	**11,80**	**12,02**	**10,09**
NB-Whisper	21,79	17,77	16,26	19,21	16,17
Google USM	24,14	28,37	15,54	17,17	18,23
NB-Wav2Vec2	24,87	22,71	21,54	21,48	19,22
Microsoft Azure	24,35	25,97	24,40	24,46	24,42
Open AI Whisper	31,73	29,66	30,80	30,05	27,27
Google Cloud long	33,61	32,11	28,83	28,43	27,07
NB-Whisper Semantic	34,30	27,71	27,91	33,07	32,08

https://ai.nb.no
https://huggingface.co/NbAiLab



National Library of Norway

NB-Whisper Large

Audio | YouTube

YouTube or Twitter/X URL

https://www.youtube.com/watch?v=_uv74o8hG30

Output Language

(•) Bokmål ( ) Nynorsk ( ) English

[x] Return timestamps

Advanced Options ◄

Clear | Submit

Examples

YouTube or Twitter/X URL	Output Language	Return timestamps
https://www.youtube.com/watch?v=_uv74o8hG30	Bokmål	Verbatim
https://www.youtube.com/watch?v=YcBWSBRuk0Q	Bokmål	Verbatim
https://www.youtube.com/watch?v=vauTloX4HkU	Bokmål	Verbatim
https://www.youtube.com/watch?v=WHF74ppqKFQ	Bokmål	Verbatim
https://www.youtube.com/watch?v=b8nz4sh_sj4	Bokmål	Verbatim
https://www.youtube.com/watch?v=vauTloX4HkU	Bokmål	Verbatim

Video



Transcription

[00:00:05.960 -> 00:00:11.760] På låven sitter nissen med sin julegrøt, så god og søt, så god og søt.
[00:00:11.840 -> 00:00:17.800] Han nikker og han smiler og han er så glad, for julegrøten vil han gjerne ha.
[00:00:17.880 -> 00:00:23.750] Men rundt omkring står alle de små rotter, og de skotter, og de skotter.
[00:00:23.830 -> 00:00:30.070] De vil så gjerne ha litt julegodter Og de danser, danser rundt i ring, ring, ring.
[00:00:30.150 -> 00:00:36.070] Men nissen, se han truer med sin store skje Nei, bare se, og kom av sted.
[00:00:36.150 -> 00:00:46.170] For julegrøten min, den vil jeg ha i fred. Og ingen, ingen vil jeg dele med. Men råtner de hopper, og de danser, og de svinser, og de svanser.
[00:00:46.250 -> 00:00:49.450] De klorer etter grøten, og de stanser.

Per Ivar Lied, Sp


STORTINGET
10:10:40

Per Ivar Lied, Sp

TINY

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferg, og også for markedsadgangen for norsk fisk.

SMALL

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferd, og også for markedsadgangen for norsk fisk.

LARGE

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferd, og også for markedsadgangen for norsk fisk.

TINY

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferg, og også for markedsadgangen for norsk fisk.

SMALL

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferd, og også for markedsadgangen for norsk fisk.

LARGE

Helsekontroll er også viktig for å forebygge spredning av sykdommer og sikre god fiskevelferd, og også for markedsadgangen for norsk fisk.

Per Ivar Lied, Sp

TINY

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk en merking av fiskeprodukt.

SMALL

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk, en merking av fiskeprodukt.

LARGE

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk enn merking av fiskeprodukter.

TINY

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk en merking av fiskeprodukt.

SMALL

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk, en merking av fiskeprodukt.

LARGE

Senterpartiet mener det vil være mer effektivt å styrke regelverket og tilsyn med helse, velferd og kvalitet av oppdrettsfisk enn merking av fiskeprodukter.


STORTINGET
10:10:40

Runar Sjåstad, Ap



Komiteen noterer seg at statsråden i ovennevnte brev mener intensjonen i forslagene allerede er ivaretatt i dagens regelverk.

Runar Sjåstad, Ap

TINY

Under komiteens behandling av representantforslaget har det kommet inn fire høringsinnspill fra henholdsvis teknad for brukerrådet Dyrevernsaliansen og

SMALL

Under komiteens behandling av representantforslaget har det kommet inn fire høringsinnspill fra forbrukerrådet, Dyrevernsalliansen og NOA.

LARGE

Under komiteens behandling av representantforslaget har det kommet inn fire høringsinnspill fra henholdsvis Tekna, Forbrukerrådet, Dyrevernsalliansen og NOAH.

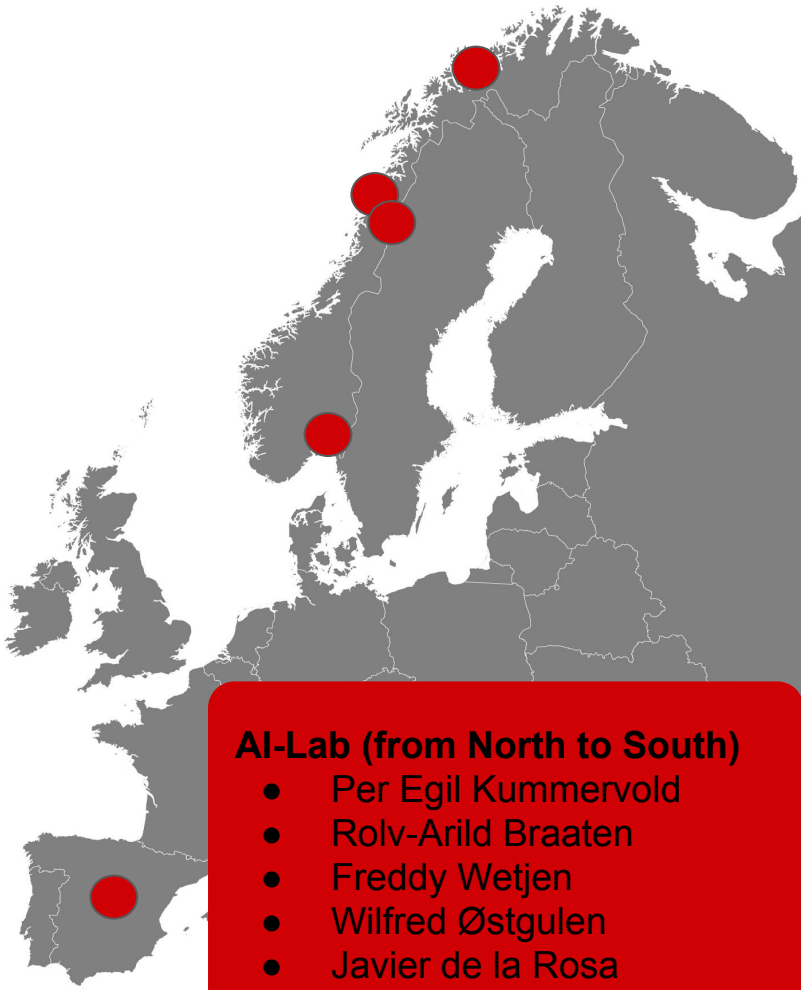
Olav Grotle, H - NB-Whisper Tiny

Olav Grotle, H - NB-Whisper Large



President, helsekontroll er et viktig element i arbeidet med å forebygge spredning av sykdommer og sikre god fiskevelferd.

https://ai.nb.no



AI-Lab (from North to South)

- Per Egil Kummervold
- Rolv-Arild Braaten
- Freddy Wetjen
- Wilfred Østgulen
- Javier de la Rosa

Google Research

Thank you!

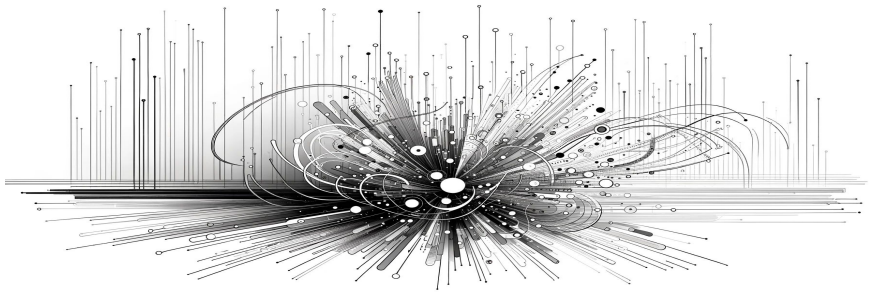
Questions?

Per Egil Kummervold
per.kummervold@nb.no



AI-lab
National Library of Norway

Basic Encoder Idea

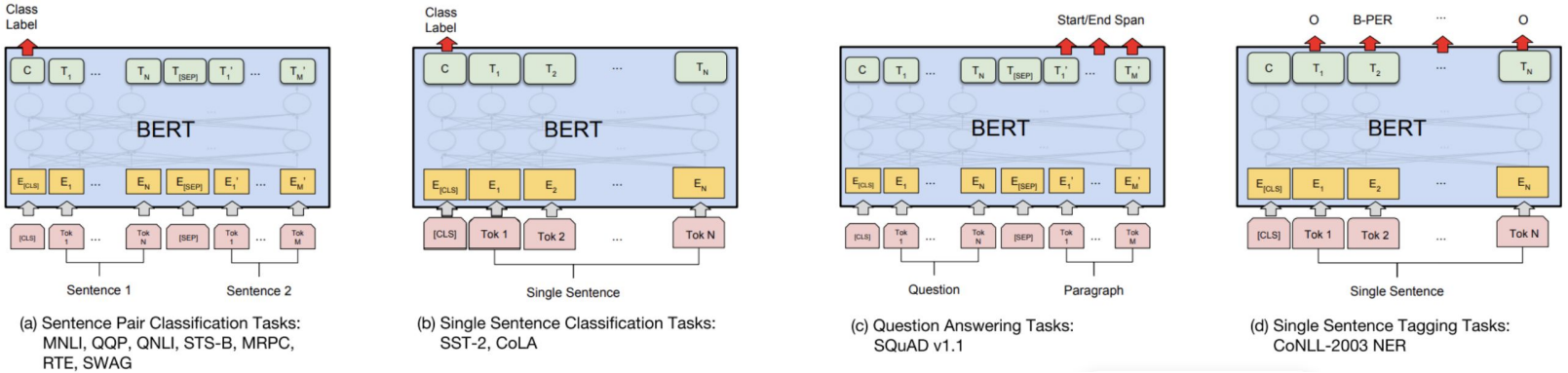


TEXT

- Classification
- Search
- Sentiment Analysis
- NER/POS



Finetuning an Encoder-model



- When you have a pretrained feature extractor, all you need to do is:
  - Add another full connected layer
  - Fit to a small finetuning dataset