CS 224S / Linguist 285
Spoken Language Processing

Tolúlọpẹ́ Ògúnrẹ̀mí | Stanford University | Spring 2025

Lecture 11: State-of-the-art deep learning approaches for speech recognition

Outline

- Conformer Architecture
- Whisper Architecture

Conformer

Conformer: Convolution-augmented Transformer for Speech Recognition

- Sequence-to-sequence transformer with multi-headed self attention.
- Combines attention (global context) with convolution (local invariance)
- RNN-T loss architecture

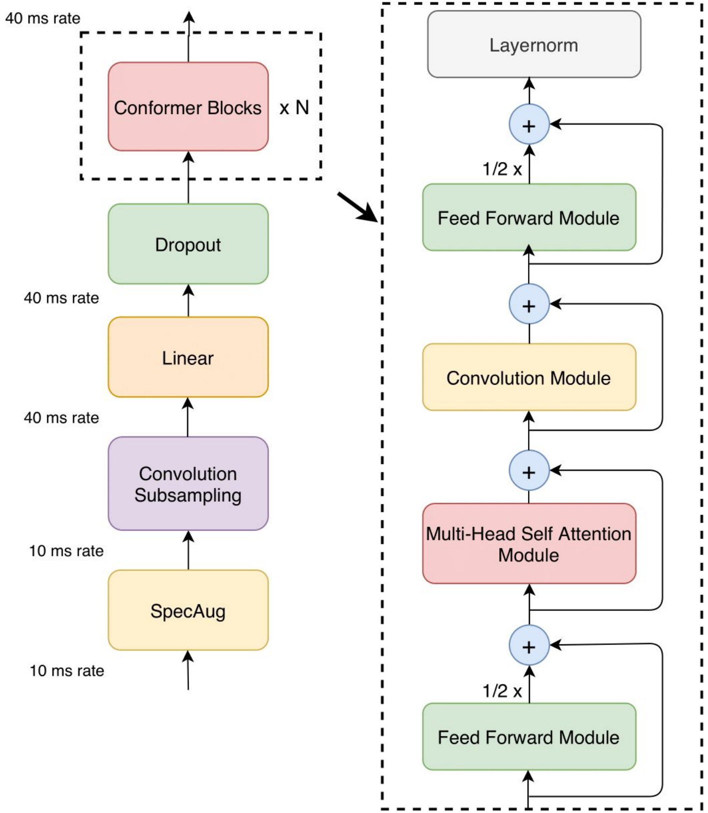
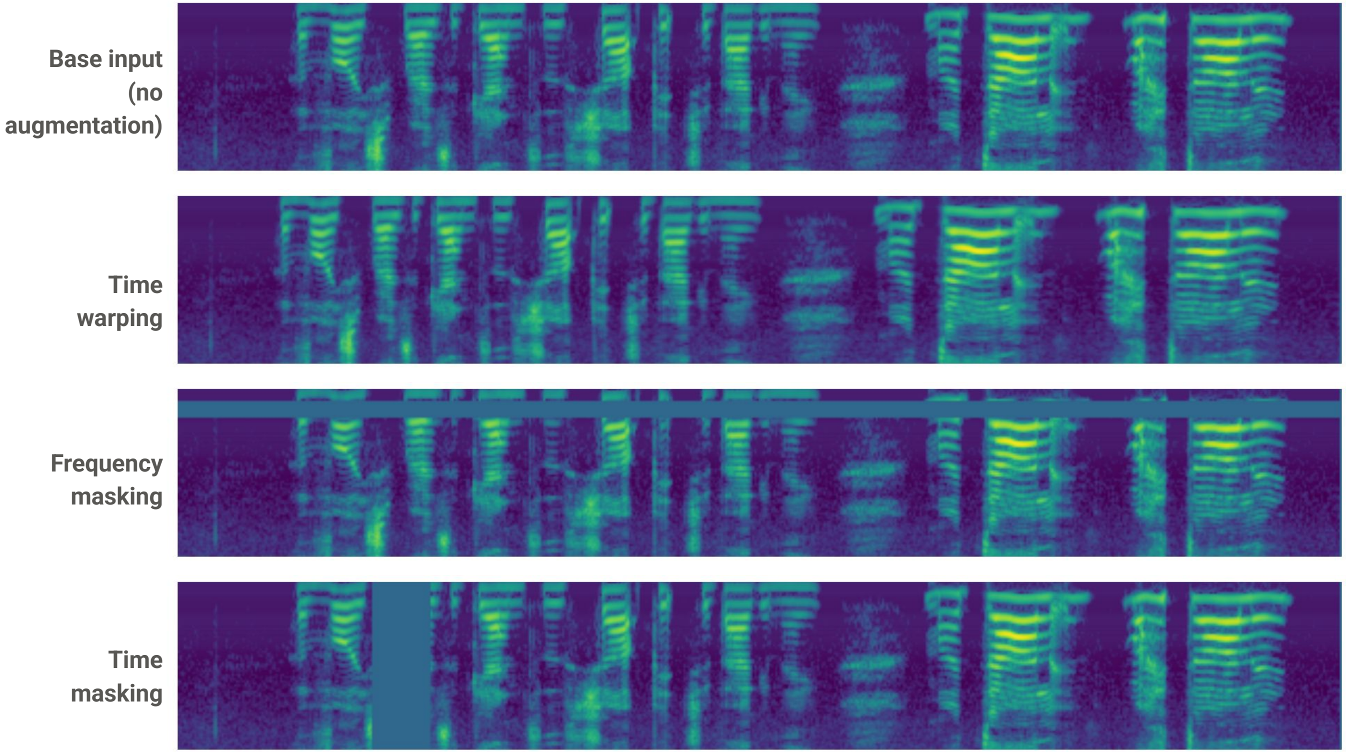
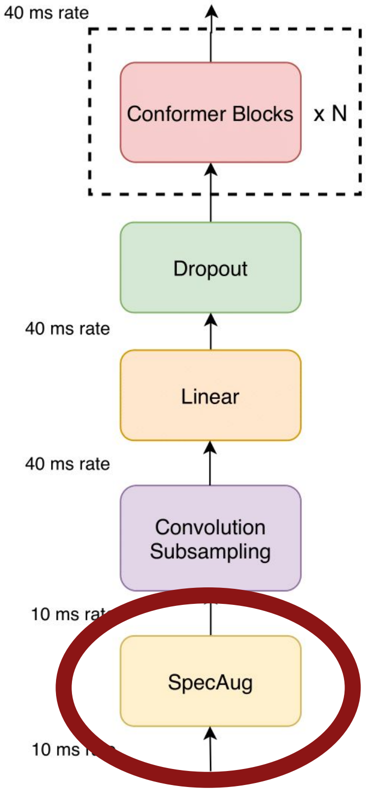


Figure: Conformer encoder model architecture. Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.

(Gulati *et al.* 2020)

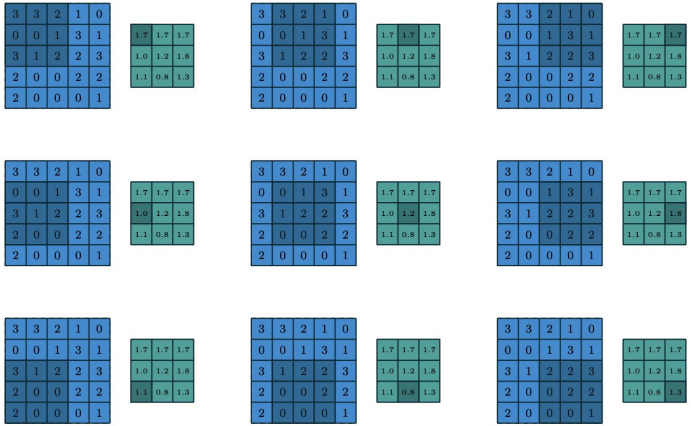
SpecAugment: Augmenting data to increase dataset size





Park *et al.* 2019

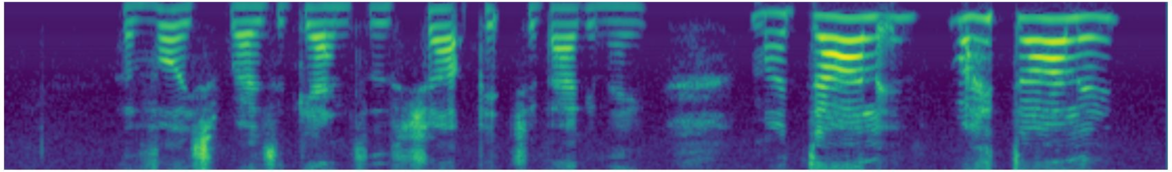
Convolution subsampling: Log mel spectrogram is an image

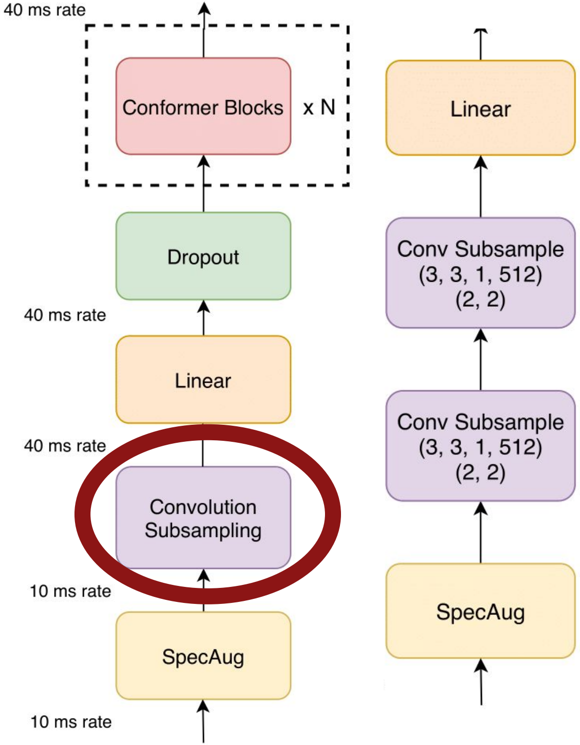


3	3	2	1	0
0	0	1	3	1
3	1	2	2	3
2	0	0	2	2
2	0	0	0	1

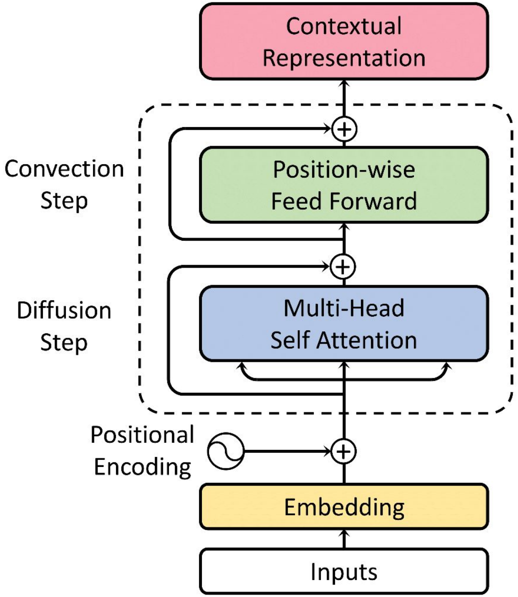
1.7	1.7	1.7
1.0	1.2	1.8
1.1	0.8	1.3

Conv image source





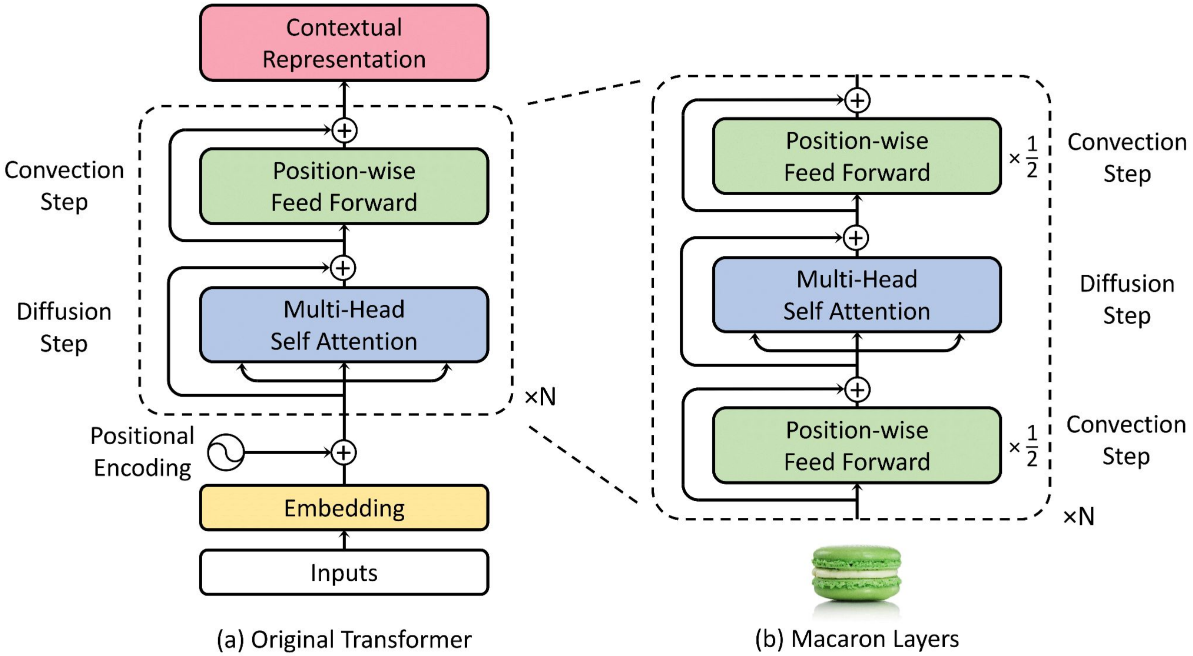
Transformer encoder to conformer encoder


Contextual Representation
Convection Step
Position-wise Feed Forward
Diffusion Step
Multi-Head Self Attention
Positional Encoding
Embedding
Inputs

(a) Original Transformer

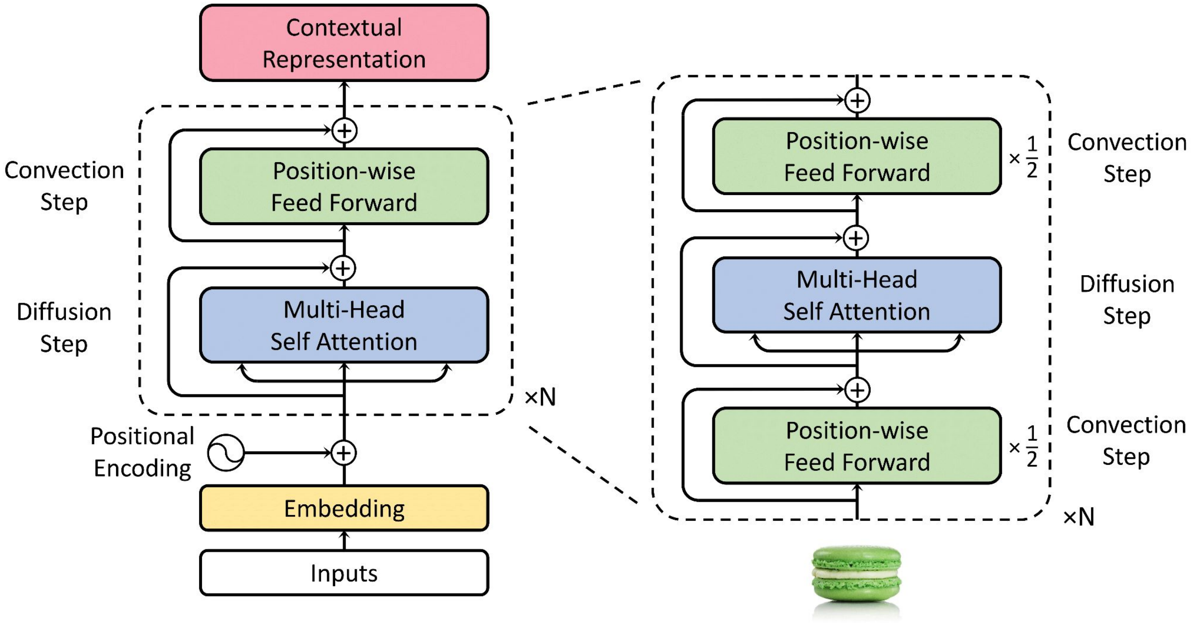
“Macaron” transformer layers from Macacron Net (Lu et. al, 2019)

Transformer encoder to conformer encoder



"Macaron" transformer layers from Macacron Net (Lu et. al, 2019)

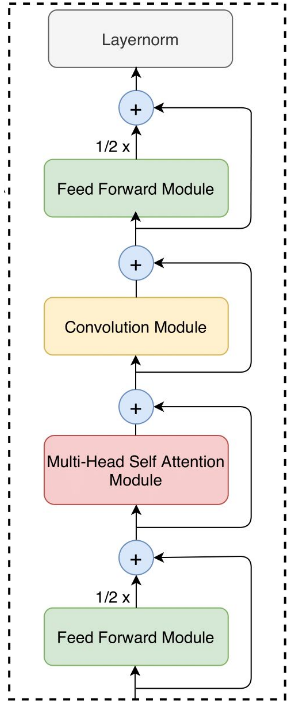
Transformer encoder to conformer encoder


Contextual Representation
Convection Step
Position-wise Feed Forward
Diffusion Step
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Position-wise Feed Forward
× 1/2
Convection Step
Multi-Head Self Attention
Diffusion Step
Position-wise Feed Forward
× 1/2
Convection Step
×N

(a) Original Transformer

(b) Macaron Layers

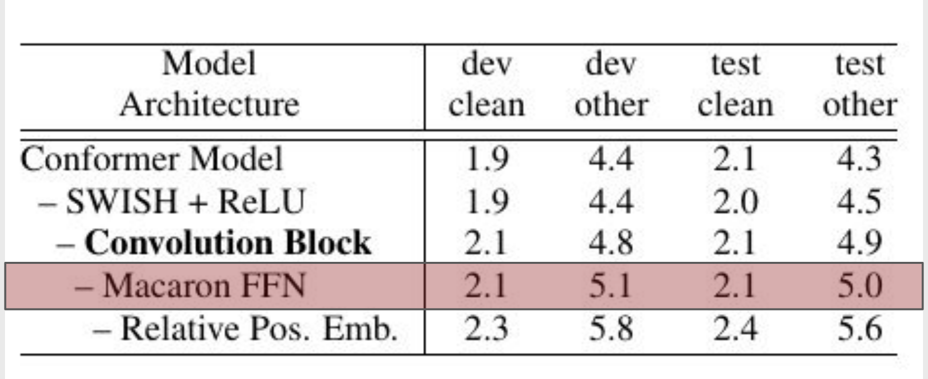
Add a Convolution Module



“Macaron” transformer layers from Macacron Net (Lu et. al, 2019)

Macaron feed-forward has a measurable effect on performance

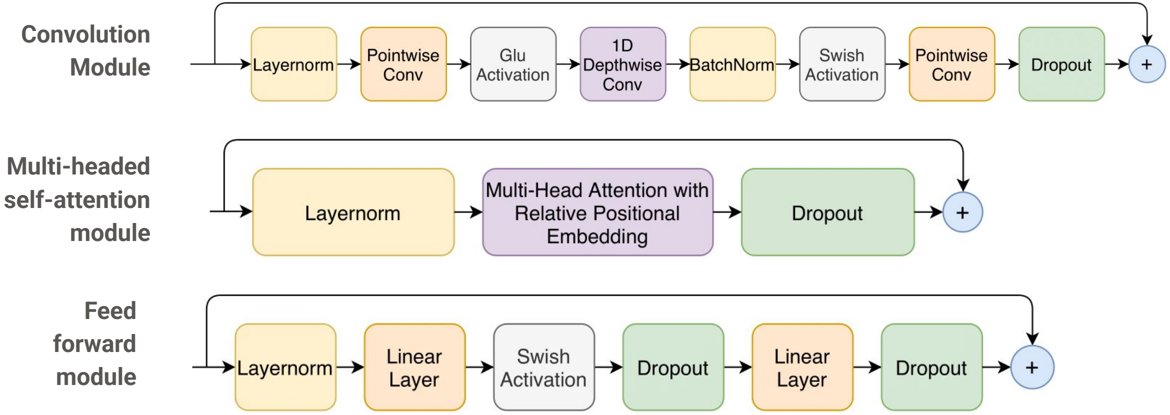
- From Ablation experiment of paper, removing unique features and moving towards a vanilla Transformer block
- All ablation study results are evaluated without the external LM.

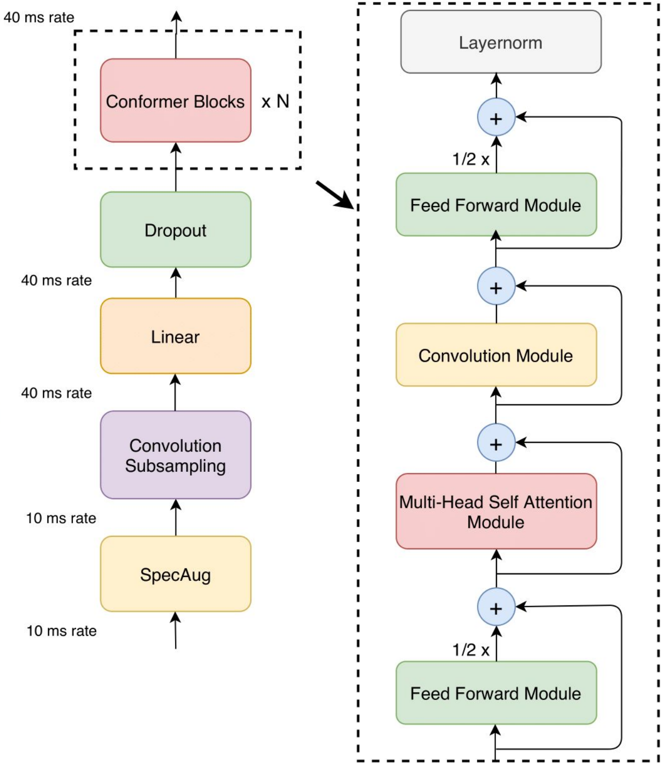


Model Architecture	dev clean	dev other	test clean	test other
Conformer Model	1.9	4.4	2.1	4.3
– SWISH + ReLU	1.9	4.4	2.0	4.5
– **Convolution Block**	2.1	4.8	2.1	4.9
– Macaron FFN	2.1	5.1	2.1	5.0
– Relative Pos. Emb.	2.3	5.8	2.4	5.6

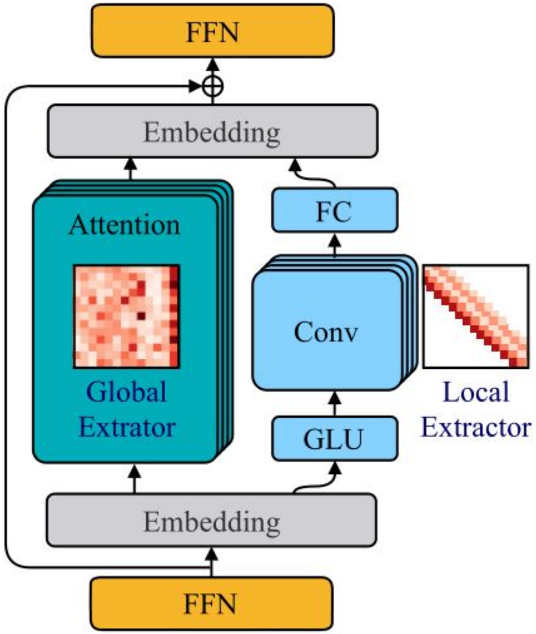
Conformer: Convolution-augmented Transformer for Speech Recognition

- Sequence-to-sequence transformer with multi-headed self attention. Directly optimizes target word sequence
- Combines attention (global context) with convolution (local invariance)

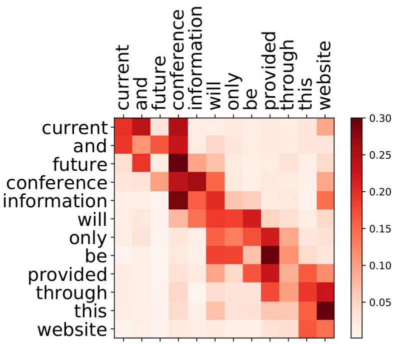




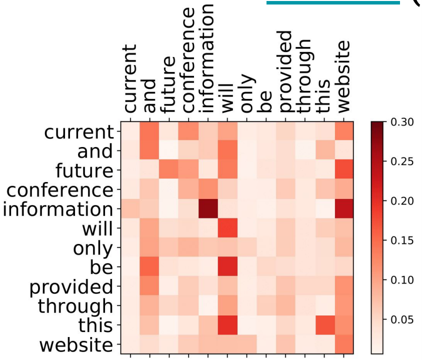
Isolating local context: Convolution module



Conventional Attention



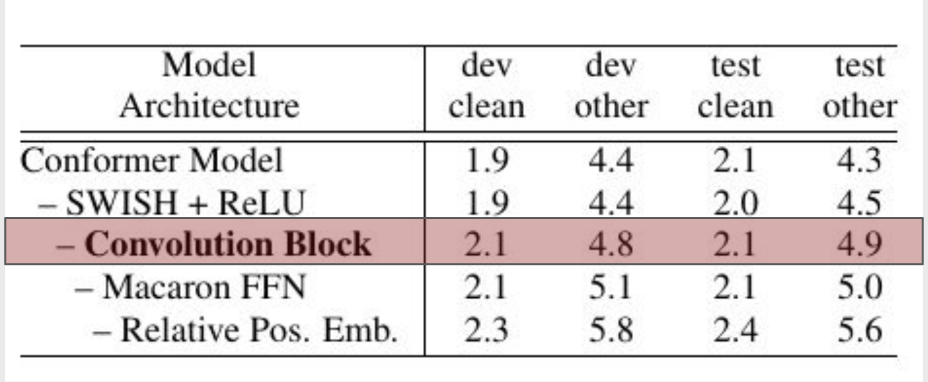
LRSA Attention Wu et al. (2020)



- Wu et al. (2020) introduce a Lite Transformer which uses a convolutional branch in a transformer layer to specialise in local feature extraction to allow the attention module to focus on global context
- They test their Lite Transformer on Machine Translation, Abstractive Summarisation and Language modelling.

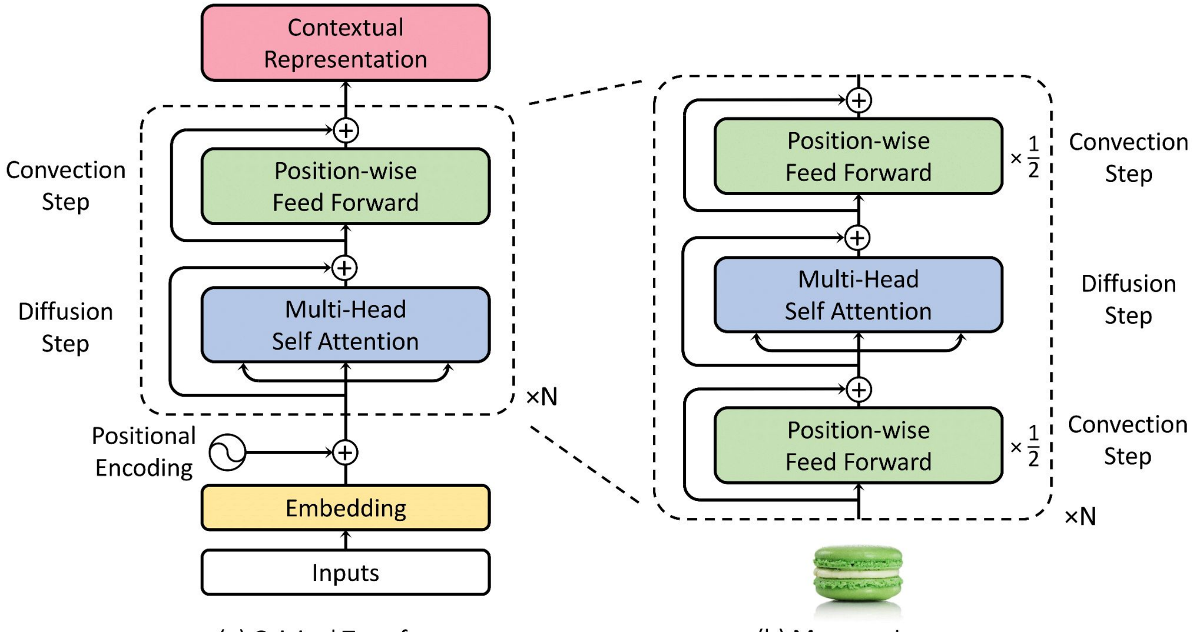
The convolution block has a the greatest effect on performance

- From Ablation experiment of paper, removing unique features and moving towards a vanilla Transformer block
- All ablation study results are evaluated without the external LM.



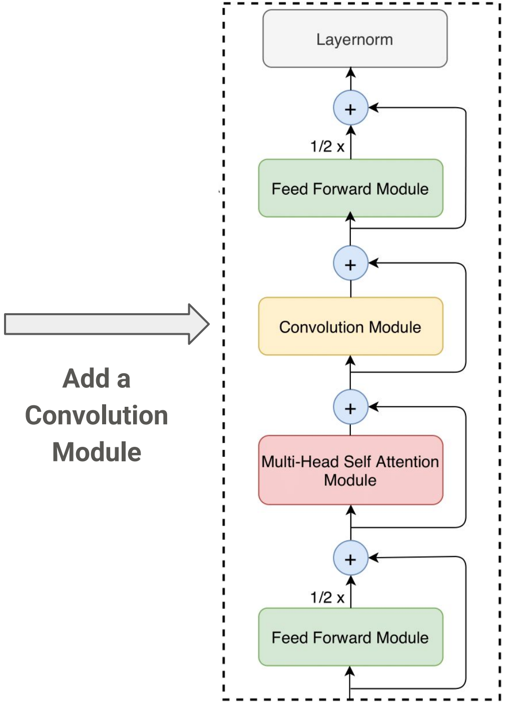
Model Architecture	dev clean	dev other	test clean	test other
Conformer Model	1.9	4.4	2.1	4.3
– SWISH + ReLU	1.9	4.4	2.0	4.5
– Convolution Block	2.1	4.8	2.1	4.9
– Macaron FFN	2.1	5.1	2.1	5.0
– Relative Pos. Emb.	2.3	5.8	2.4	5.6

A conformer is a transformer encoder with a convolutional module in the transformer layer


Contextual Representation
Convection Step
Position-wise Feed Forward
Diffusion Step
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Position-wise Feed Forward
× 1/2
Convection Step
Multi-Head Self Attention
Diffusion Step
Position-wise Feed Forward
× 1/2
Convection Step
×N

(a) Original Transformer

(b) Macaron Layers

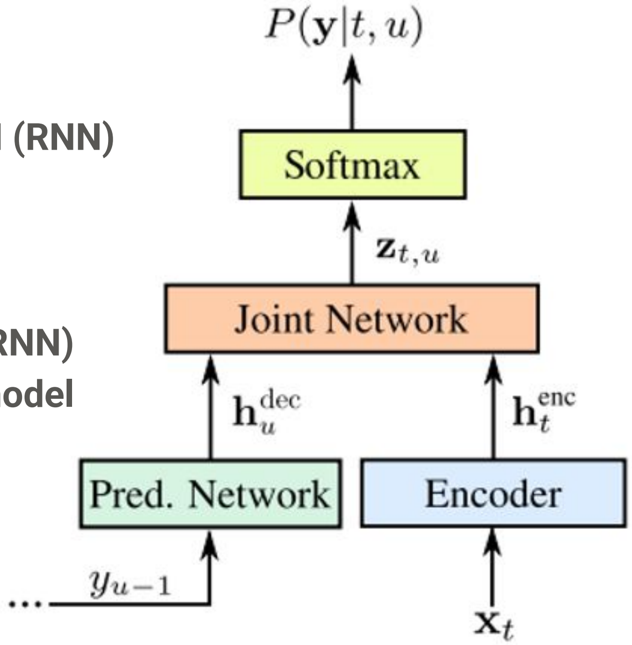


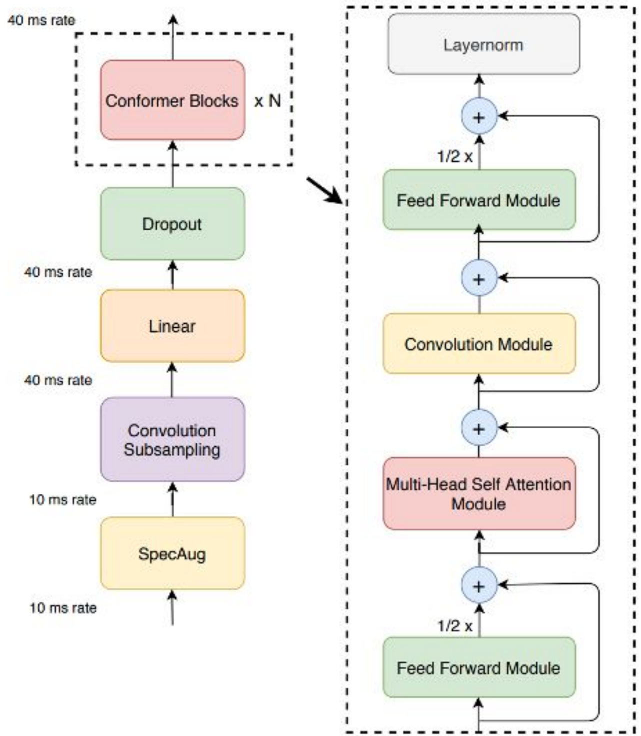
“Macaron” transformer layers from Macacron Net (Lu et. al, 2019)

Conformer decoder is an LSTM

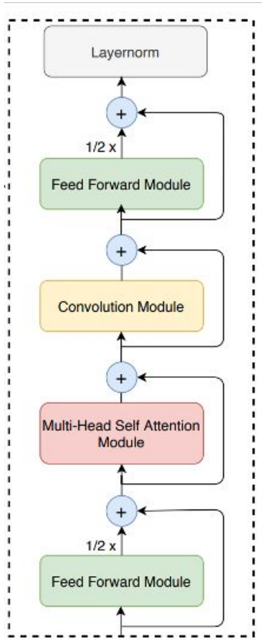
1 hidden layer LSTM (RNN)

3 hidden layer LSTM (RNN)
Pre-trained as language model

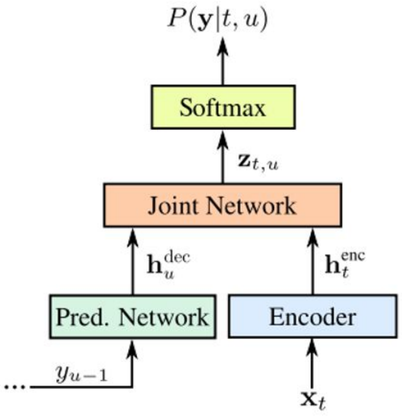




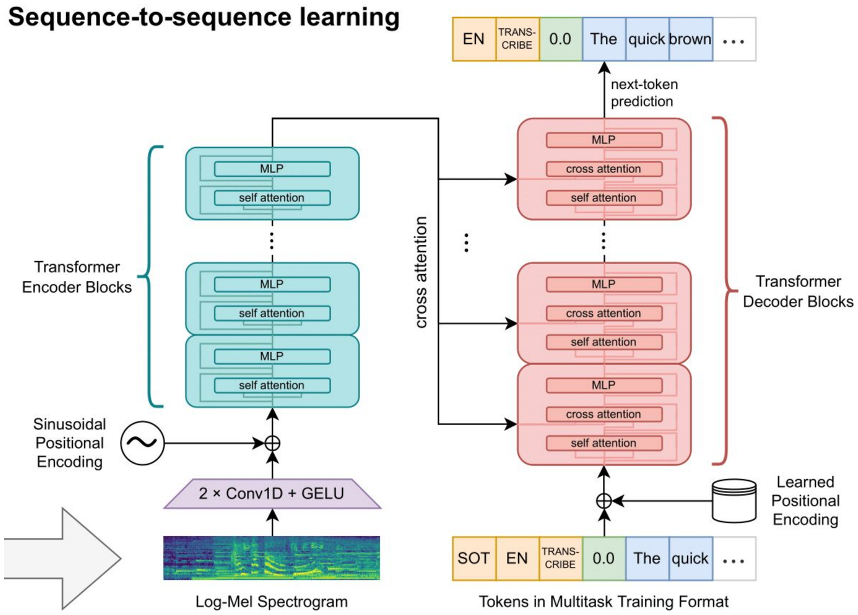
Whisper decoder is a transformer decoder



Conformer encoder



Conformer decoder



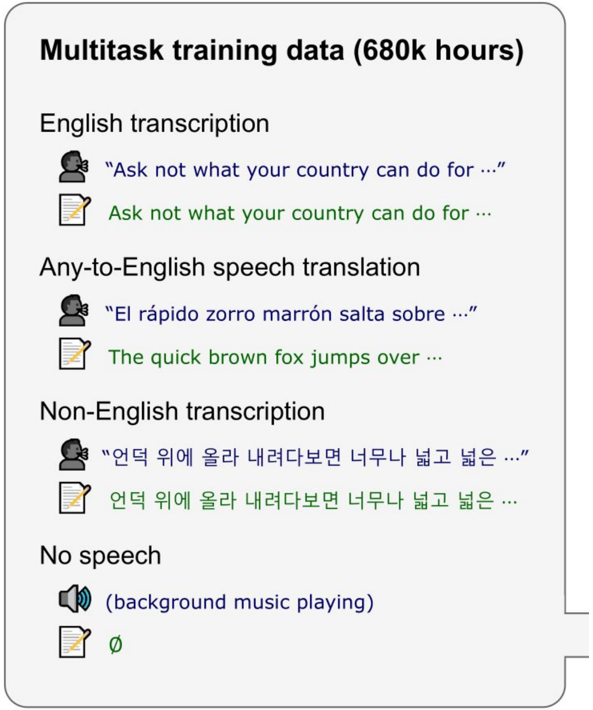
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
...
Tokens in Multitask Training Format

Whisper encoder

Whisper decoder

Whisper

What can whisper do?



Multitask training data (680k hours)

English transcription

🗣 "Ask not what your country can do for …"

📝 Ask not what your country can do for …

Any-to-English speech translation

🗣 "El rápido zorro marrón salta sobre …"

📝 The quick brown fox jumps over …

Non-English transcription

🗣 "언덕 위에 올라 내려다보면 너무나 넓고 넓은 …"

📝 언덕 위에 올라 내려다보면 너무나 넓고 넓은 …

No speech

🔊 (background music playing)

📝 ∅

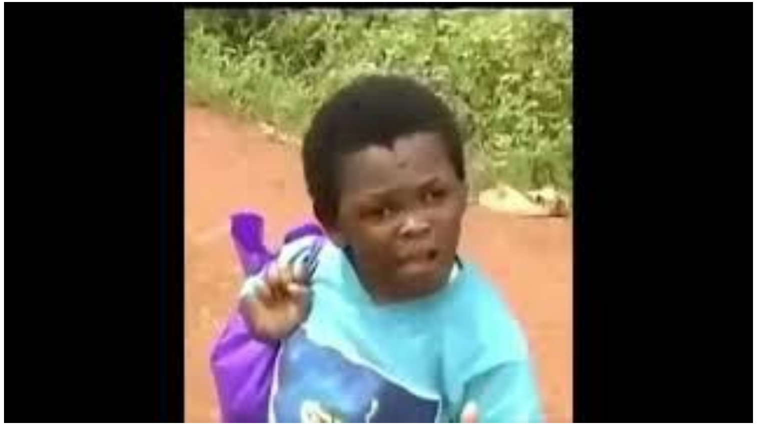
(Radford et al. 2022)

Ground truth: Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

Whisper (French): Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

Whisper (Translation): Like me, you may have already experienced this scene. What are you doing this summer? I don't know yet

What can Whisper do?



Nigerian Accented English

Ground truth: Please, please, I don't have any time for any gossip now. Ehn? Yes

Whisper: Please, please, I don't have any time for any gossip now. Yes.

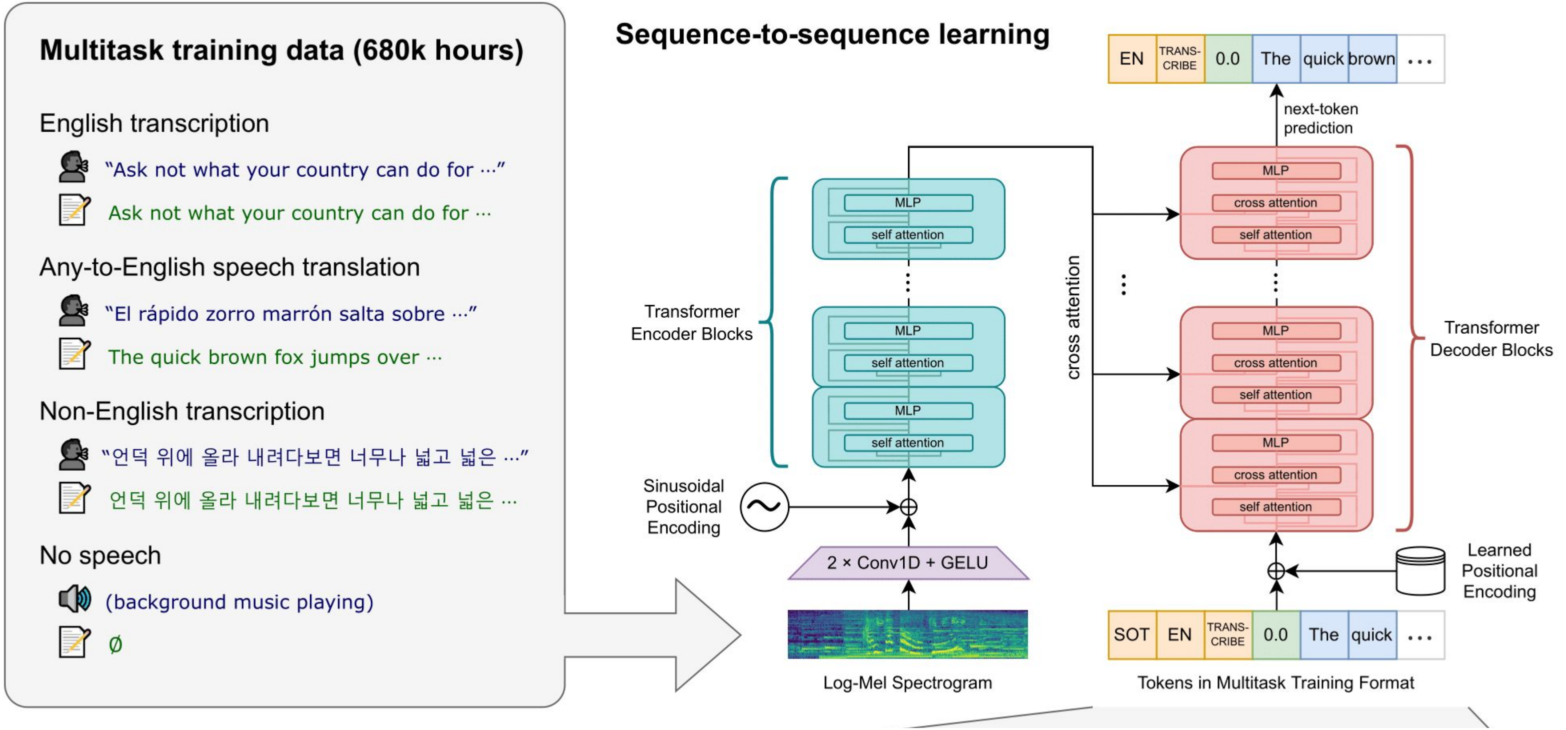
Swahili

Ground truth: Jambo! Jambo! Jina lako ni nani? Jina langu ni Juma. Unatoka wapi? Ninatoka Australia…..

Whisper: Jambo! Jambo! Jinalako ni nani? Jinalangu ni Juma. Unatoka wapi? Ninatoka Australia…..

Whisper (Translation): Hello! Hello! What is your name? My name is Juma. Where are you from? I am from Australia.

How does Whisper perform a variety of tasks accurately?



Multitask training data (680k hours)

English transcription

"Ask not what your country can do for ⋯"

Ask not what your country can do for ⋯

Any-to-English speech translation

"El rápido zorro marrón salta sobre ⋯"

The quick brown fox jumps over ⋯

Non-English transcription

"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"

언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

(background music playing)

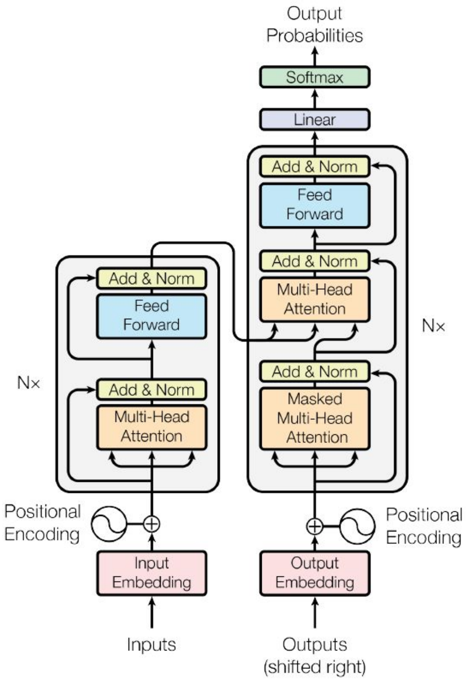
∅

Sequence-to-sequence learning

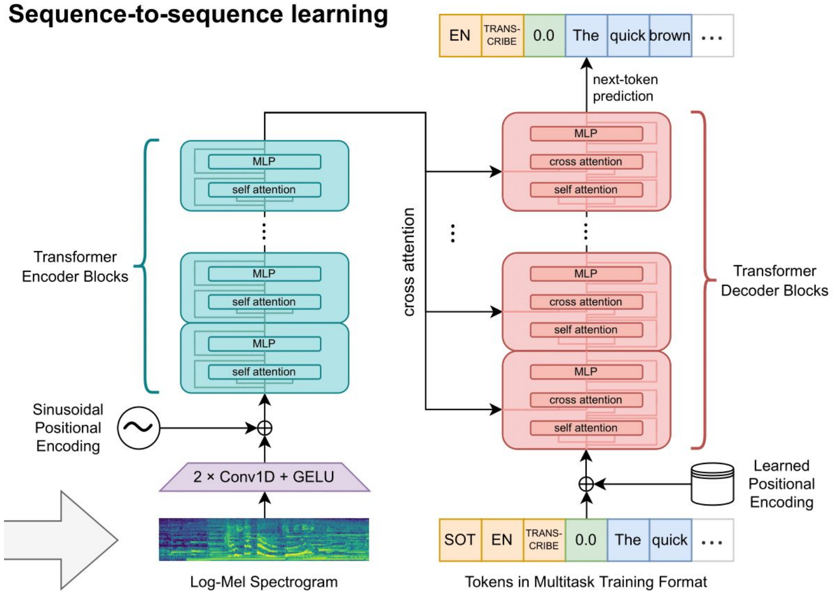
(Radford et al. 2022)

Whisper is a encoder-decoder transformer model

Text encoder-decoder transformer

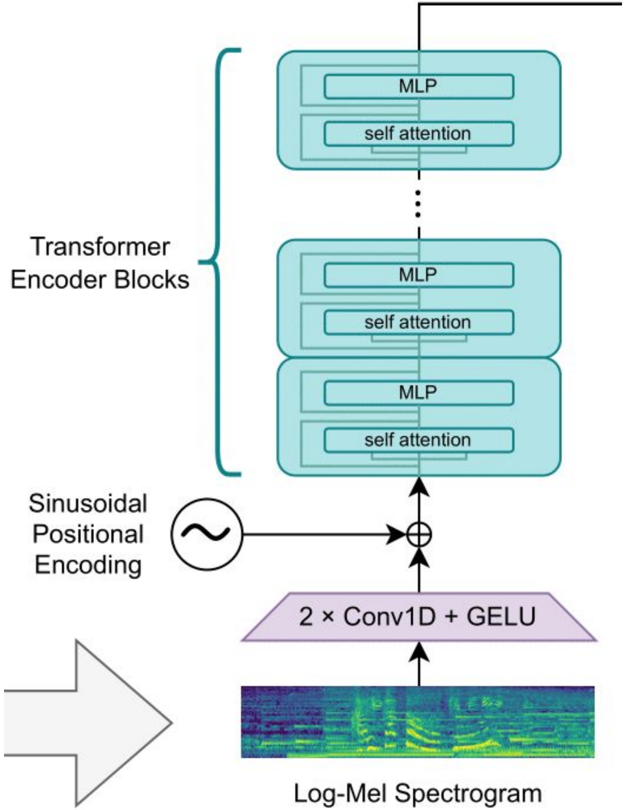


Whisper encoder-decoder transformer



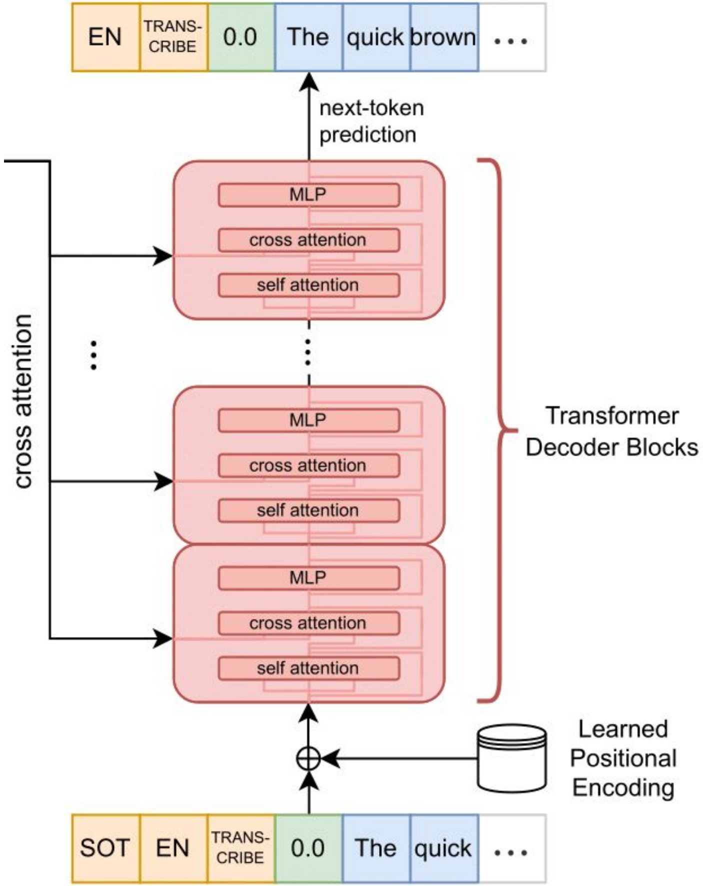
Whisper encoder: transformer encoder

- Encoder input: log-mel spectrogram and two convolutional layers
- Positional embeddings
- Standard transformer blocks depending on model size

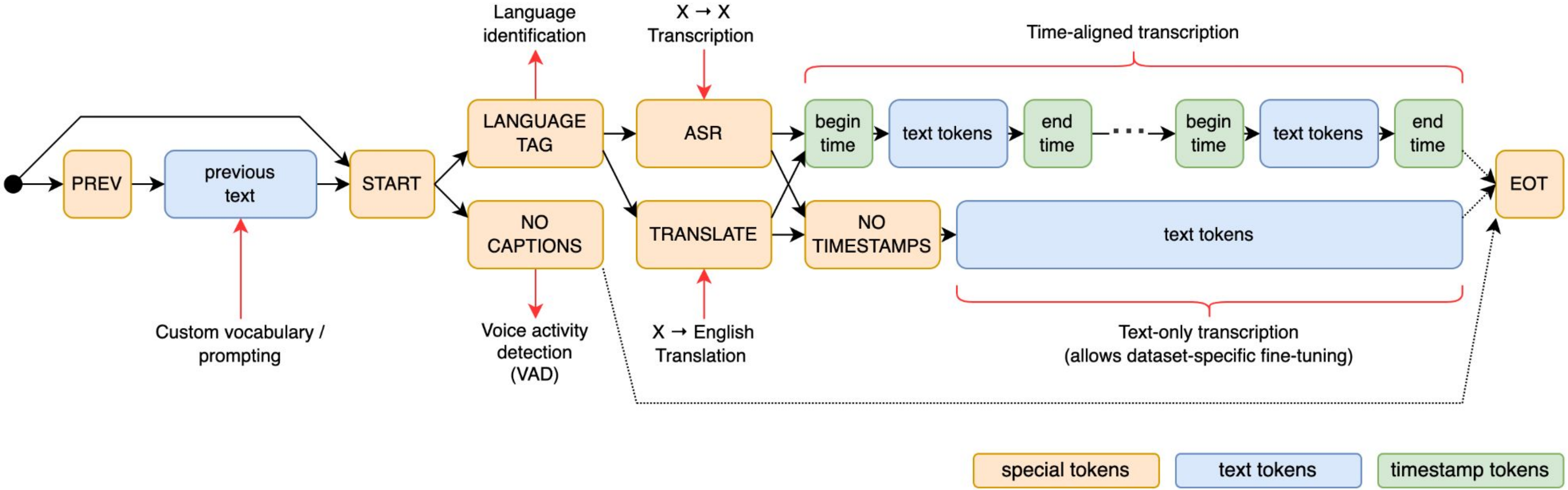


Whisper decoder: autoregressive transformer decoder

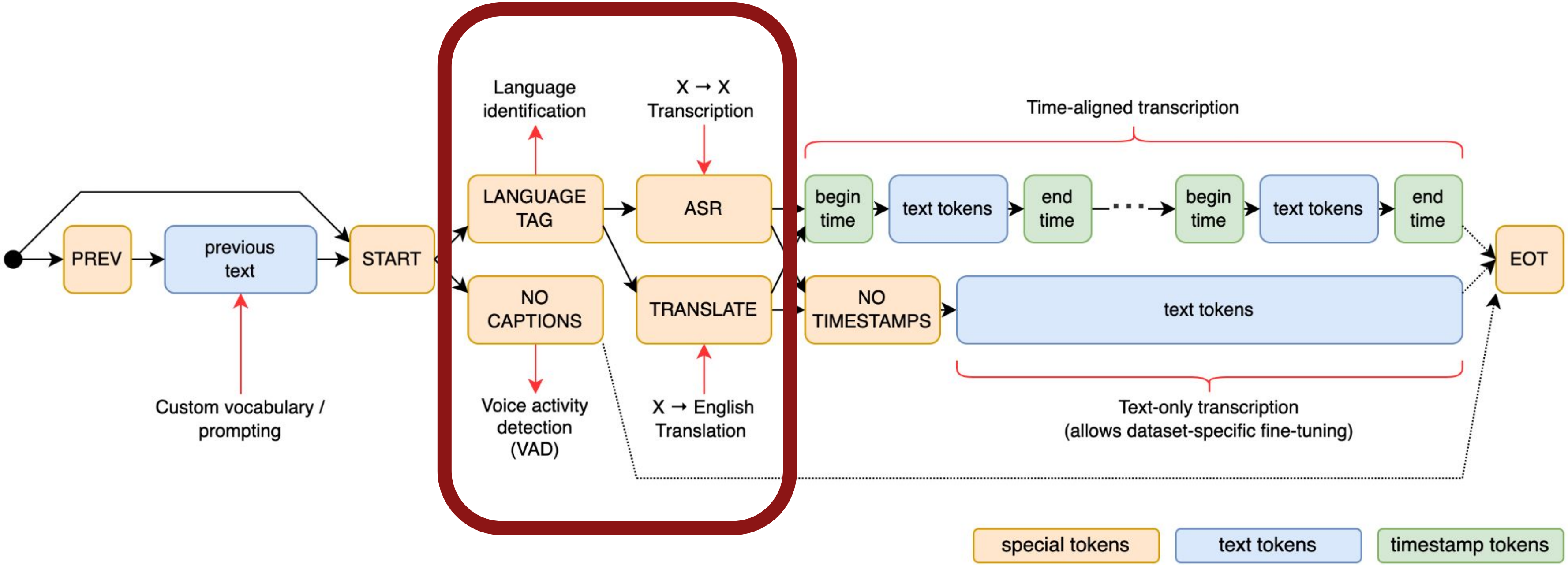
- GPT-2 style model - no more CTC loss
- Custom vocabulary and tags:
  - Language tags: <en>, <fr>
  - Task tags: <transcribe>, <translate>
  - Timestamps
  - Speech tags: <nospeech>



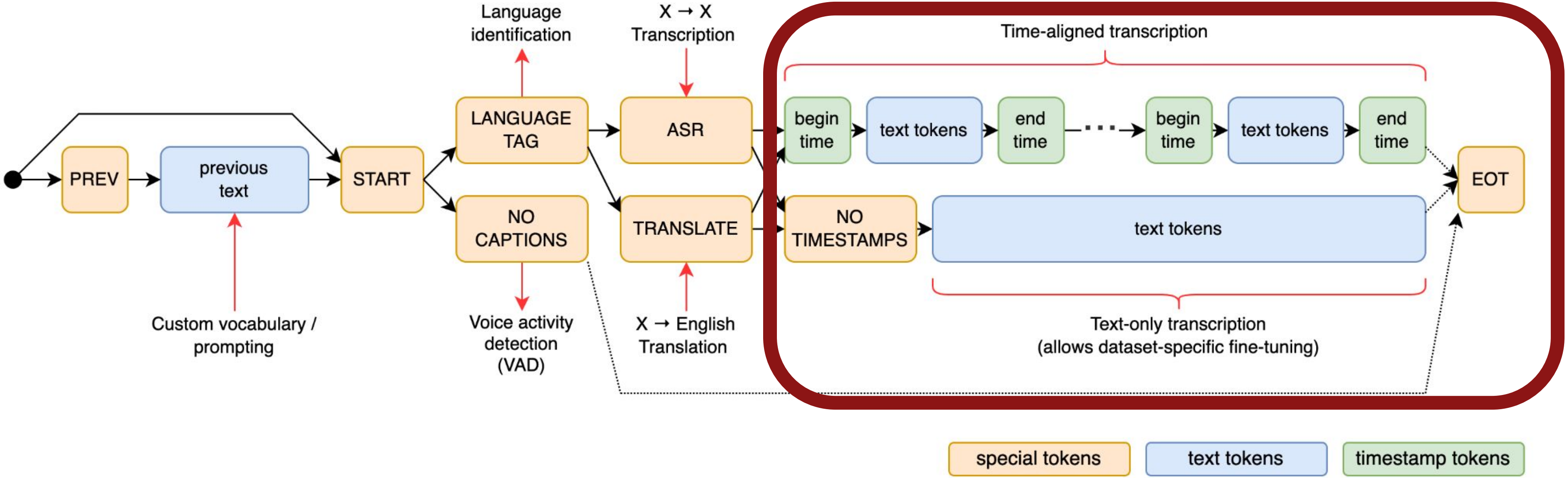
Whisper (Multitask Training)



Whisper: task tags



Whisper: time stamps and transcription tokens



Whisper output example

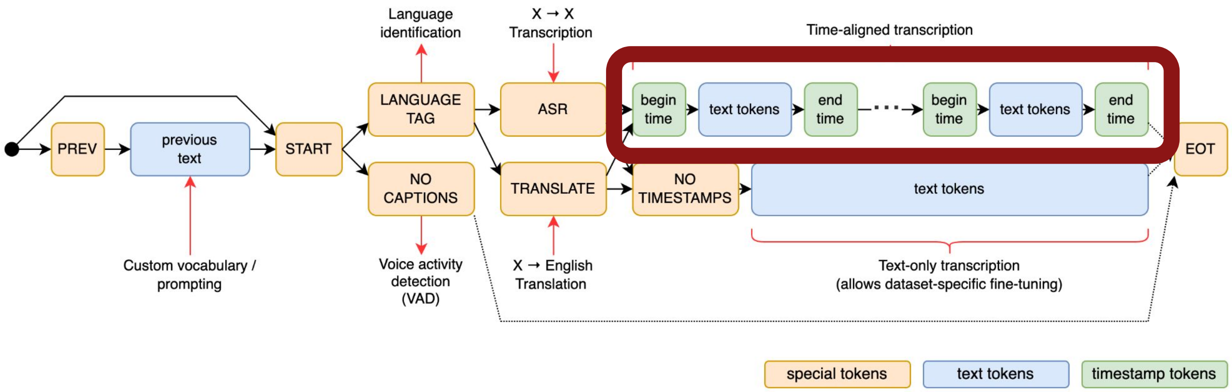
Ground truth: Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

Whisper (French): ['<|startoftranscript|>', '<|fr|>', '<|transcribe|>', '<|notimestamps|>', ' Comme', ' moi', ',', ' vous', ' avez', ' peut', '-', 'être', ' déjà', ' v', 'éc', 'u', ' cette', ' scène', '.', ' Qu', "'", 'est', '-', 'ce', ' que', ' tu', ' fais', ' cet', ' été', ' toi', ' ?', ' É', 'c', 'oute', ',', ' je', ' ne', ' sais', ' pas', ' encore', ' trop', ……………. '<|endoftext|>']]

Whisper (Translation): ['<|startoftranscript|>', '<|fr|>', '<|translate|>', '<|notimestamps|>', ' Like', ' me', ',', ' you', ' may', ' have', ' already', ' experienced', ' this', ' scene', '.', ' What', ' are', ' you', ' doing', ' this', ' summer', '?', ' I', ' don', "'t", ' know', ' yet', ………. '<|endoftext|>']

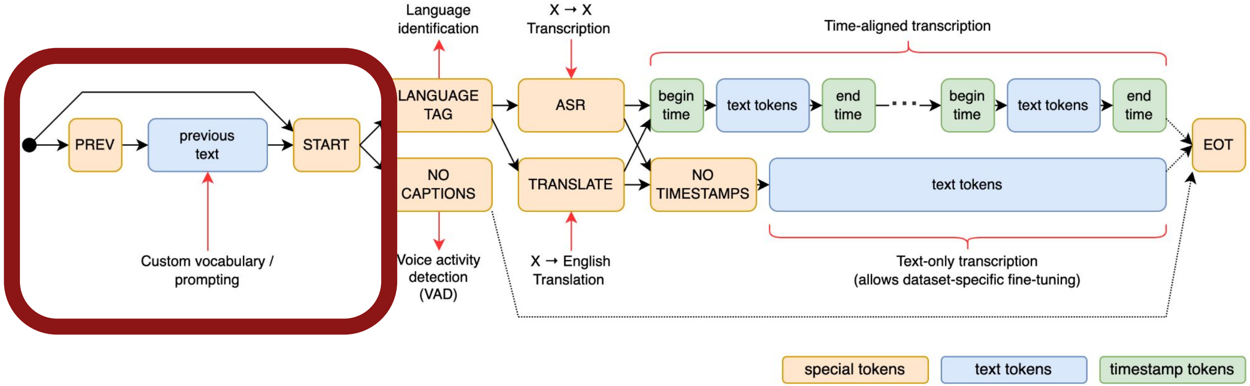
Whisper handles long-form transcription

- Whisper only takes 30 seconds of speech at a time
- Timestamp tokens are used to determine when to shift the window



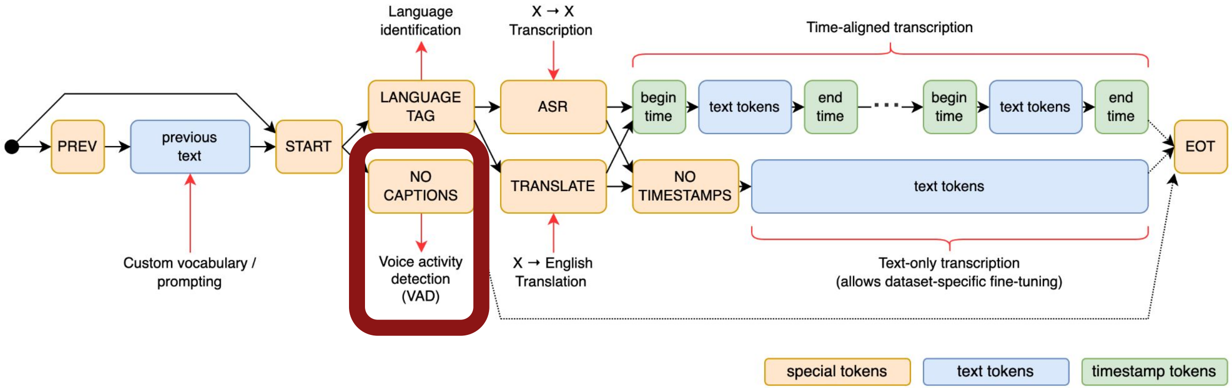
Whisper handles long-form transcription

- Whisper only takes 30 seconds of speech at a time
- Timestamp tokens are used to determine when to shift the window
- Text from the previous window is provided with transcribing current window



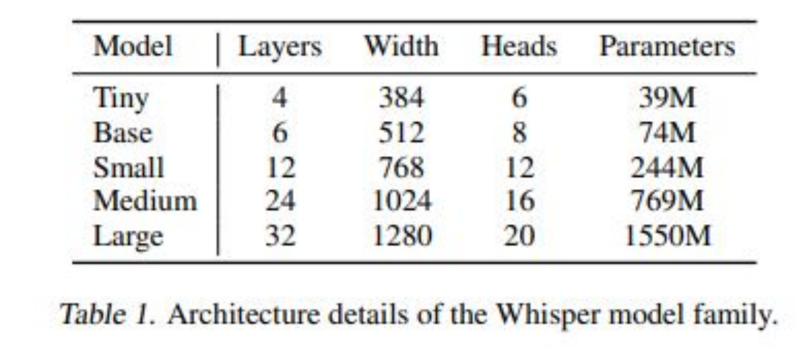
Whisper handles long-form transcription

- Whisper only takes 30 seconds of speech at a time
- Timestamp tokens are used to determine when to shift the window
- Text from the previous window is provided with transcribing current window
- Voice activity detection using <|nospeech|> tag



Model sizes and related trade-offs

- Variety of model sizes trades of accuracy, compute needs, and model size
  - Whisper tiny can run locally on modern smartphones!
  - Whisper large is competitive with fairly costly, complex cloud ASR systems
- Training and fine tuning
  - Ability to fine tune smaller models to specialize for tasks/languages can yield great performance with Tiny - Small sizes
  - Ability to distill (e.g. student/teacher training) to improve small models given large model teacher signals
- Whisper vs straight ASR systems
  - Whisper great for prototyping due to flexibility. Possibly conformer / ASR-specific models are better in practice if your task is monolingual ASR only
  - Whisper is over-represented in online discussion about ASR tools due to public availability and ease of use



Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 1. Architecture details of the Whisper model family.

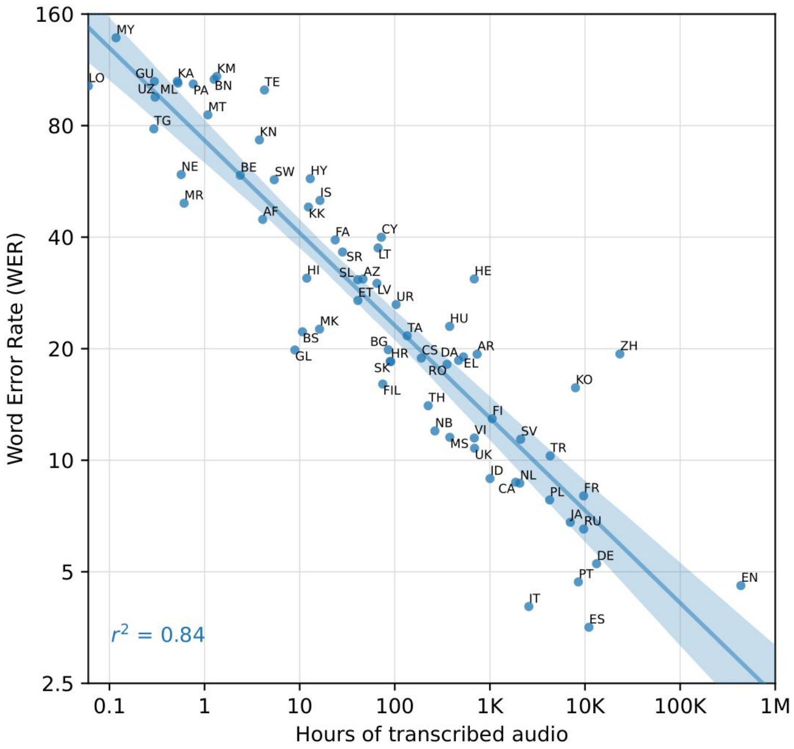
Challenging cases and LLM-like failure modes

Ground truth: “Allez, allez, allez, alleeeeez….. Martinez! Qui vient de sauver…. Face à Randal Kolo Muani. Mais c’est fou ! Et ça repart de l'autre côté, et ça repart de l'autre côté !

Whisper (French): Allez, allez.....

Whisper (Translation): Come on, come on,..............

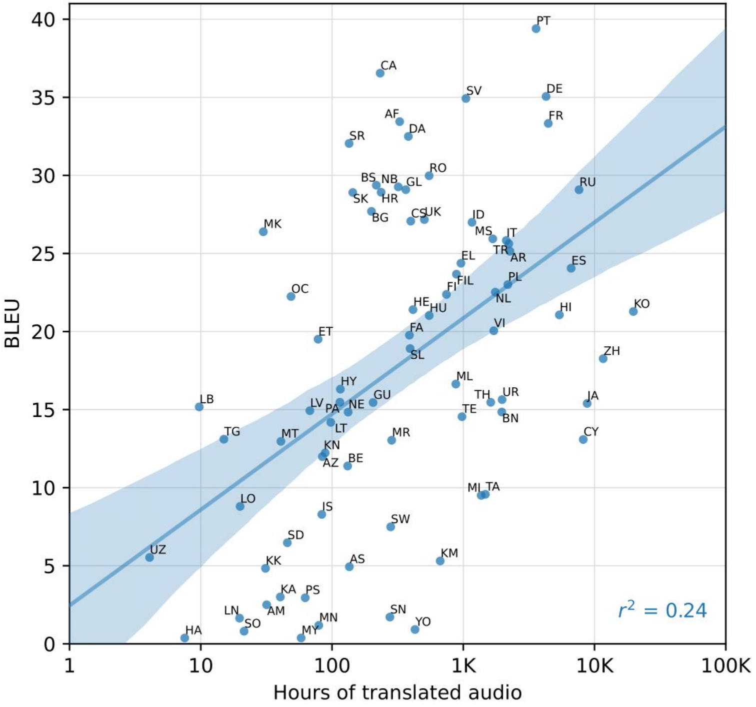
More training data, better ASR performance



Transcription performance on FLEURS dataset (lower is better)

- Whisper transcribes text in different languages with varied performance
- We get a WER below 10 once we pass 1000 hours of training data
- Low resource languages (top left) have a WER above 100
- Transcription performance correlates well with training data

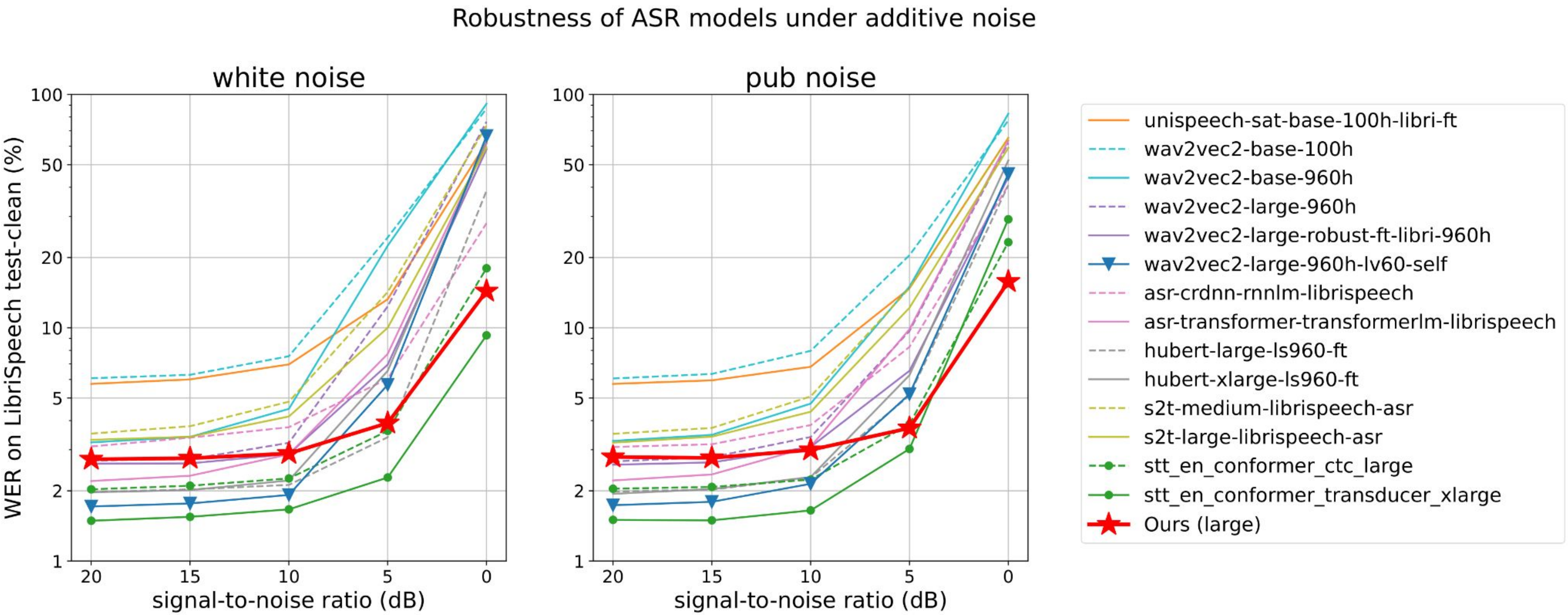
Whisper’s automatic speech translation performance



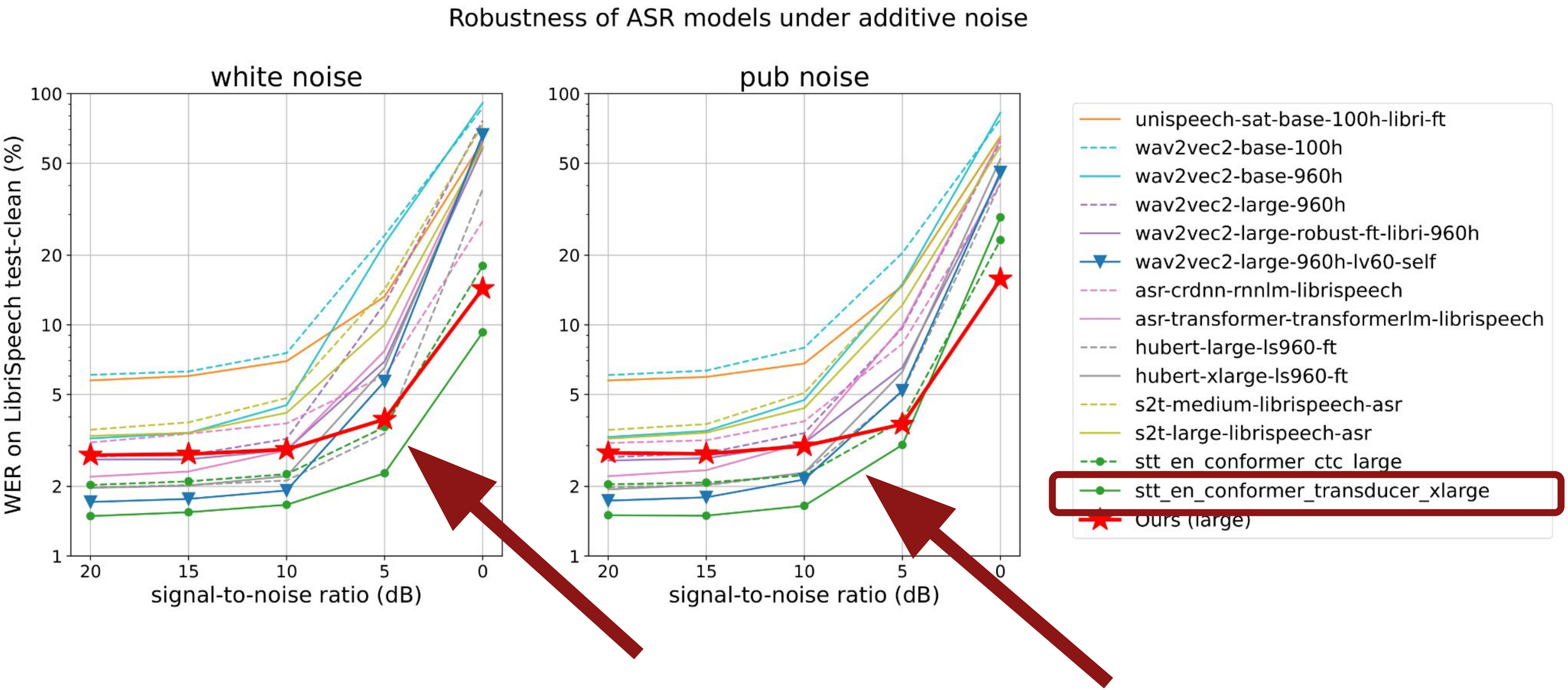
Speech translation performance on FLEURS dataset (higher is better)

- Translation performance is better on average after the 100 hours of training data mark
- Amount of training data isn’t as correlated with translation performance as it is with transcription performance
- Although possible, speech translation is likely to only be reliable for high resource languages

Whisper is relatively robust to additive noise



Conformer is also robust to additive noise



What to consider when choosing an ASR solution

- **How many hours of data do you have?**
  - If you have hundreds of hours of data and need speech recognition for a specific domain, you could get away with training your own model or fine tuning an existing model
  - If you have tens of hours of data, you can finetune an existing model
- **Which language are you looking to transcribe or translate?**
  - If the model is well represented in Whisper, it is a strong baseline, especially for translation into English
  - If you'd like to translate into another language, you will need your own model
- **What is your compute budget?**
  - If you don't have any compute budget, using the best model you can find may be the best way for forward (Homework 4)
  - If you have a small compute budget, finetuning a small model variant could outperform a larger model out-of-the-box
  - If you have a large compute budget (and large enough dataset) you can consider finetuning a larger moder variant.

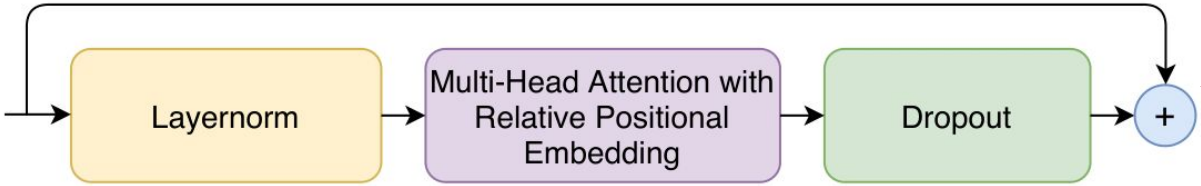
Thank You

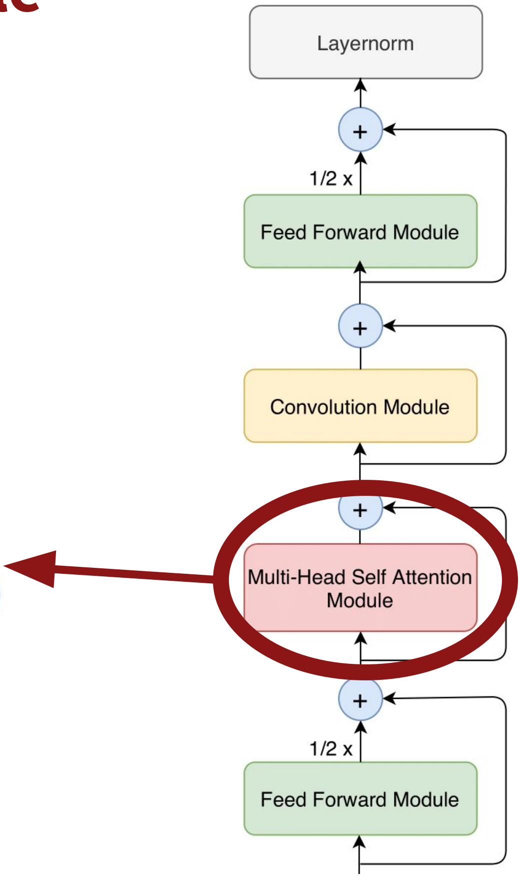
Appendix

Conformer

Global context: Multihead attention module

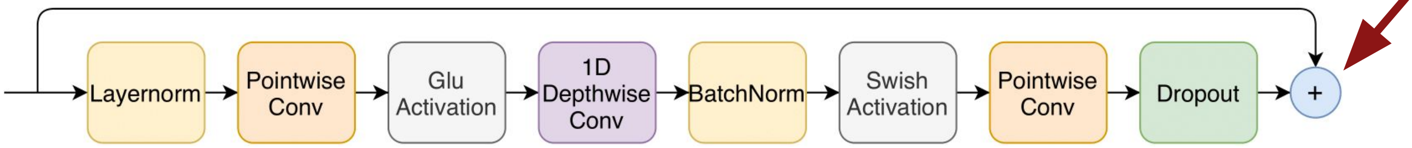
- Attention captures long range global context across the utterance (content based global interactions)
- Relative positional encoding allows the model to generalise better on different input lengths

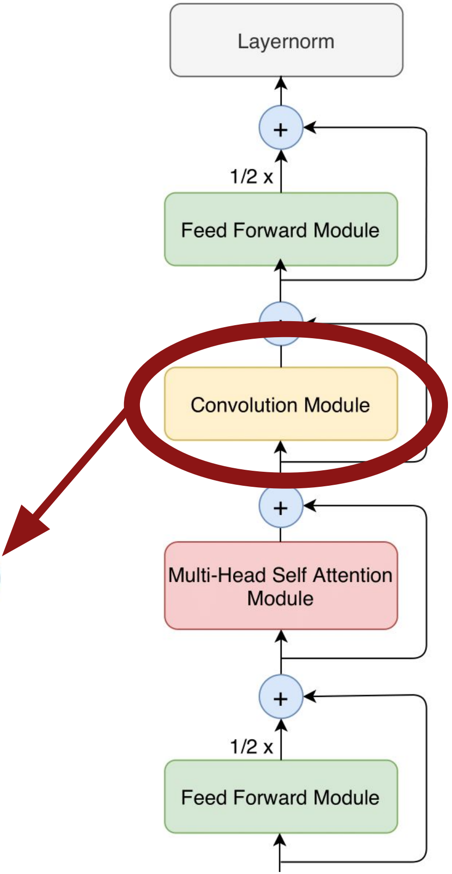




Local context: Convolution module

- Convolutions capture fine-grained local features
- Similar to Wu et al. (2020) convolutional module specialises on modelling local relationships to enable attention module to model global relationships.





Conformer: RNN-Transducer Loss

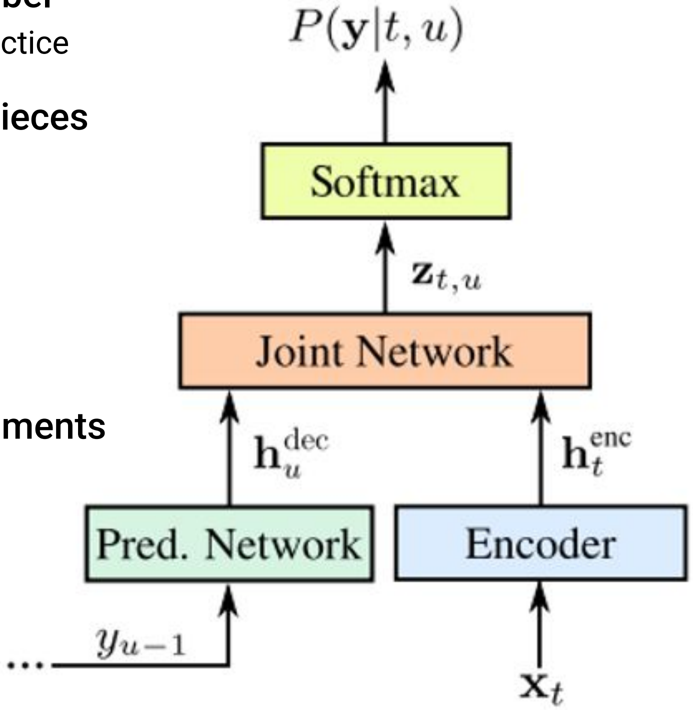
- Directly optimizes target word sequence as correct label
  - Graphemes (letters) or word parts (10k-50k) used in practice

- Learned combination of acoustic + language model pieces
- Conditions on sequence output so far (y_{t-1})
- Single alignment:

$$P(\mathbf{z}|\mathbf{x}) = \prod_i P(z_i|\mathbf{x}, t_i, \text{Labels}(z_{1:(i-1)}))$$

- Maximize P(y|x) by summing over all consistent alignments

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}(\mathbf{y},T)} P(\mathbf{z}|\mathbf{x}),$$



RNN-T loss: (Graves, 2012)
Figure: Rao, Sak, & Prabhavalkar. 2017)

Whisper

What can whisper do? (full transcripts)

Ground truth: Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

Whisper (French): Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop.

Whisper (Translation): Like me, you may have already experienced this scene. What are you doing this summer? I don't know yet

(Radford et al. 2022)

All 3 mins French

Comme moi, vous avez peut-être déjà vécu cette scène. Qu'est-ce que tu fais cet été toi? Écoute, je ne sais pas encore trop. Une discussion de bureau anodine, téléphone en main. Et toi? Camping cette année. Ah génial! Et t'as le matos, réchaud, tente et tout? J'ai rien du tout, il faut que j'achète ça. Non mais attends, achète rien, demande à Élise, elle y est allée l'année dernière, elle a tout acheté, elle a déjà tout. Ah génial, ok. Bon allez, il faut que j'y retourne. Bon à toutes, merci. Je t'en prie. A toute. Ciao. Et le lendemain, comme par hasard, une publicité parfaitement ciblée ou une nouvelle proposition d'amis. La fameuse Élise. Le micro de notre téléphone pourrait-il être utilisé à notre insu? Sommes-nous sur écoute? J'ai posé la question à Stéphanie Houillon, spécialiste en cybersécurité. Elle me propose une petite expérience. Regardez en direct l'activité de mon téléphone. A chaque fois que nous naviguons dans nos applications, notre téléphone émet un signal. Quand le micro fonctionne, ce signal augmente, car le téléphone est plus sollicité. Donc j'ai tout le trafic qui est capturé par mon ordinateur. Il voit ce qui sort et ce qui rentre de mon téléphone. Le trafic, ce sont ces lignes de code qui défilent en fonction de l'activité du téléphone. Je donne l'autorisation au micro et à la vidéo. Nous lançons une application de visioconférence qui fait donc appel à notre micro. Et là, c'est parti. Les lignes de code défilent très vite, signe d'une activité intense. Donc en effet, là, ça bouge tout le temps, il se passe des choses, on sent qu'un signal est émis régulièrement depuis mon téléphone. Oui, parce qu'il y a un flux audio et vidéo qui est actif. Maintenant, nous allons regarder ce qu'il se passe quand mon téléphone est posé à côté, comme lors de notre discussion à la machine à café. Sur l'écran, il n'y a plus de mouvement. Les lignes de code sont figées. Le micro ne semble pas enregistrer notre conversation Ça, ça nous donne la certitude qu'en ce moment a priori il n'y a pas de flux audio qui sort ou qui rentre du téléphone en continu Donc pour vous il n'y a pas d'écoute? Par rapport à ce que là nous on a observé a priori non et il n'y a pas non plus de preuves par d'autres recherches que ça serait actuellement le cas Mais il y a quand même parfois des scénarios un peu troublants Je parle de camping par exemple en vrai avec quelqu'un on a tous les deux nos téléphones dans notre poche j'ai pas fait de recherche internet là-dessus Pourtant, je reçois une pub ciblée. Comment on l'explique si ce n'est pas qu'on est sur écoute? Déjà, tu as discuté avec des collègues de bureau, avec lesquels tu es sans doute en contact sur Facebook, en tant qu'amis. Ces collègues de bureau, peut-être qu'elles, elles ont fait des recherches de camping, ou qu'elles l'ont mentionné dans des posts. La deuxième chose, c'est que oui, ces autres collègues de bureau, peut-être qu'elles en ont reparlé entre elles, par Facebook Messenger, par exemple. Comme on va utiliser plein d'applications, plein de services, les points d'entrée pour collecter des informations, ils se multiplient.

All 3 mins English

Like me, you may have already experienced this scene. What are you doing this summer? I don't know yet. A discussion of anodine office, phone in hand. Camping this year. Oh great! And you have the equipment, heater, tent and everything? I have nothing at all. I have to buy that. No, but wait, buy nothing, ask Elise, she went there last year, she bought everything, she already has everything. Oh great, okay. Okay, I have to go back there. Well, see you all, thank you. Well, I beg you. See you all. Bye. And the next day, as by chance, a perfectly targeted advertisement or a new friend proposal. The famous Elise. Could the microphone of our phone be used at our discretion? Are we listening? I asked the question to Stéphanie Houillon, a specialist in cybersecurity. She offers me a little experience. Look at the activity of my phone. Every time we navigate through our applications, our phone emits a signal. When the microphone works, this signal increases because the phone is more solicited. So I have all the traffic that is captured by my computer. So it sees what comes out and what comes in from my phone. The traffic, these are the lines of code that run according to the activity of the phone. I give authorization to the microphone and the video. We launch an application of video conferencing, which therefore calls our microphone. And there it is, it's gone. The lines of code run very fast, sign of intense activity. So, indeed, it's moving all the time, things are happening, we feel that a signal is being issued regularly from my phone. Yes, because there is an audio and video stream that is active. Now, we are going to see what happens when my phone is placed next to it, as during our discussion at the coffee machine. On the screen, there is no movement. The lines of code are frozen. The microphone does not seem to record our conversation. This gives us the certainty that at the moment, there is no audio stream that comes out or enters the phone continuously. So for you there is no listening? Compared to what we have observed, there is no and there is no proof by other research that this would be the case. But there are still sometimes a little troubling scenarios. I'm talking about camping for example, in real life with someone we both have our phones in our pockets, I didn't do any internet research on it, yet I receive a targeted ad. How do we explain it if we're not listening? Well, first you have to discuss with colleagues from the office with whom you are probably in contact on Facebook as a friend. These colleagues from the office, maybe they did research on camping or they mentioned it in posts. The second thing is that these other colleagues from offices, maybe they have spoken to each other again, through Facebook Messenger for example. Since we are going to use a lot of applications, a lot of services, the entry points to collect information, they multiply.