Dual-decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation

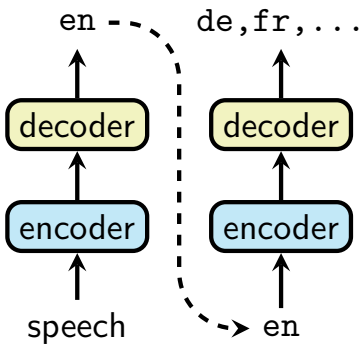
Hang Le[1] Juan Pino[2] Changhan Wang[2]
Jiatao Gu[2] Didier Schwab[1] Laurent Besacier[1]

[1]Univ. Grenoble Alpes, CNRS, LIG [2]Facebook AI

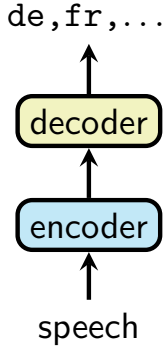
FACEBOOK AI

Existing Models

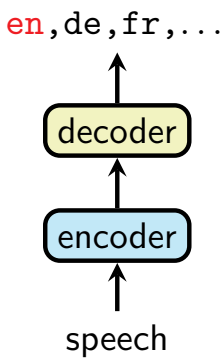
From cascade to end-to-end



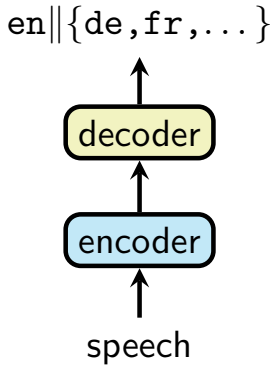
Cascade (Stentiford and Steer, 1988, Waibel et al., 1991)



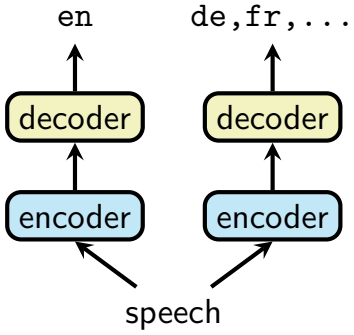
No ASR (Bérard et al., 2016; Weiss et al., 2017)



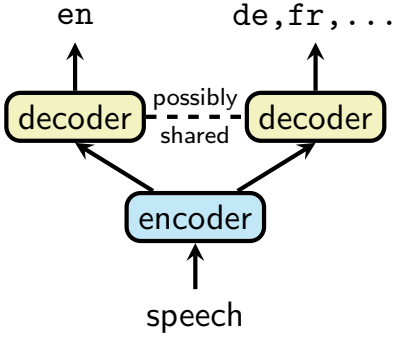
en-as-additional-language (Gangi et al., 2019)



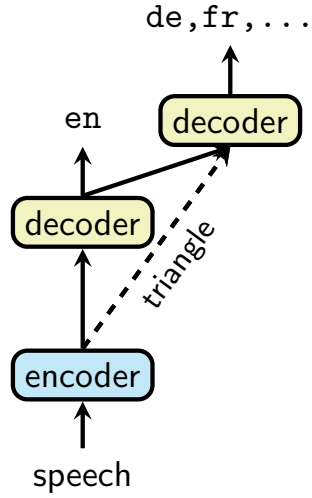
Concatenated (Sperber et al., 2020)



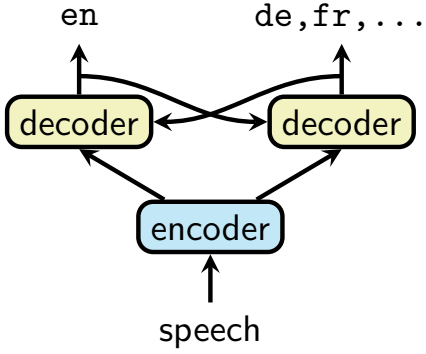
Independent (Sperber et al., 2020)



Multitask and **Shared** (Anastasopoulos and Chiang, 2018; Sperber et al., 2020)



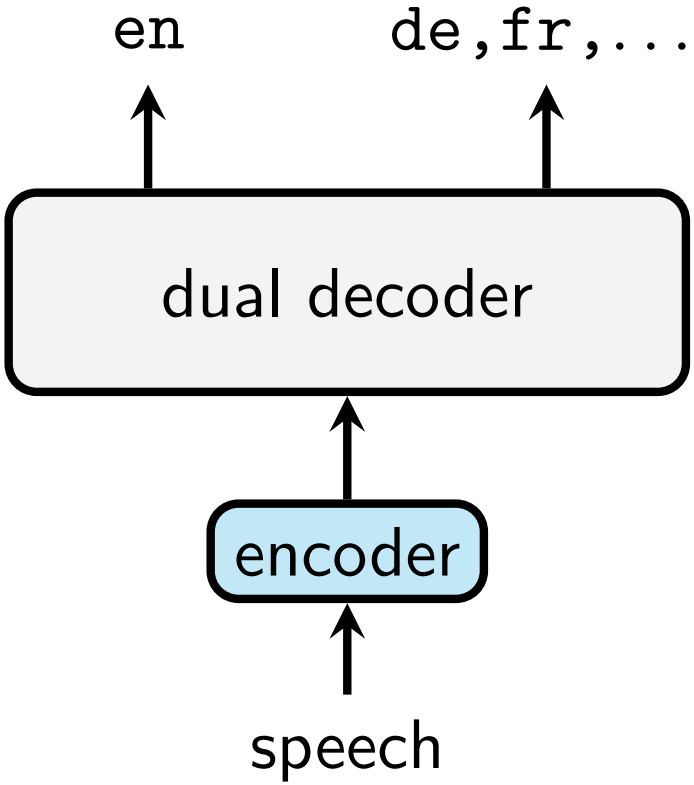
Two-stage, Triangle (Anastasopoulos and Chiang, 2018; Sperber et al., 2019)

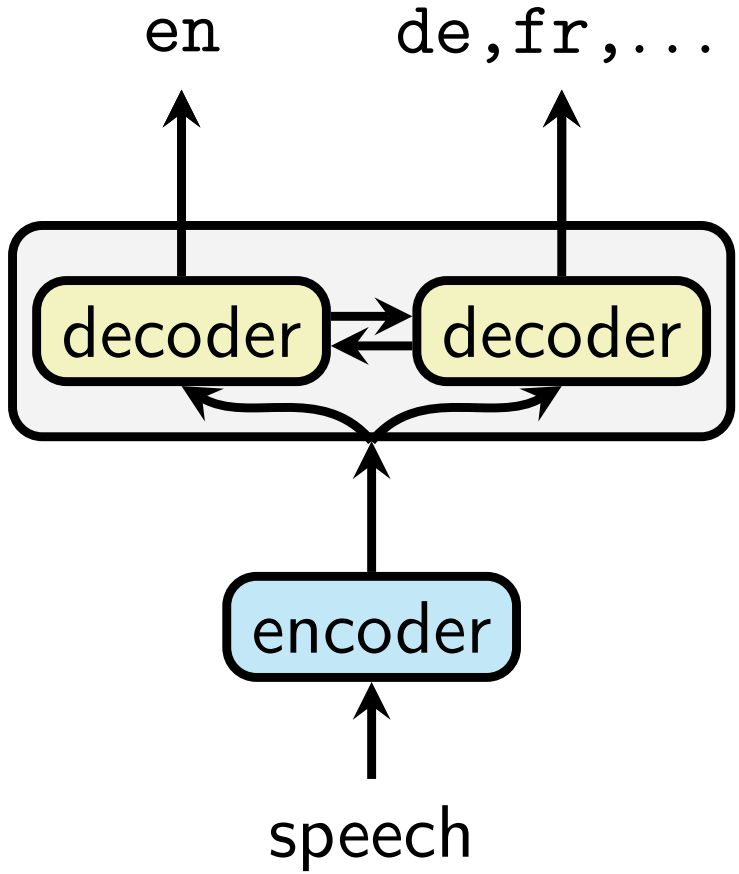


Interactive decoding (Liu et al., 2020)

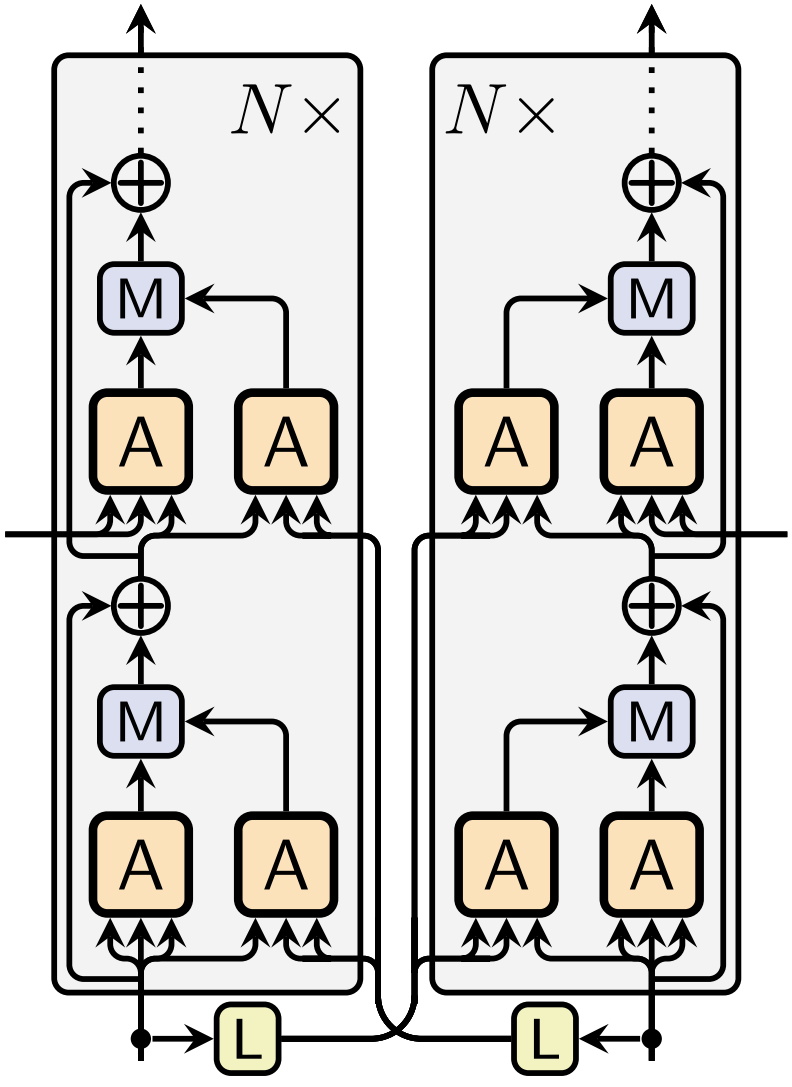
Dual-decoder Transformer

- Motivated by previous work, but *more general* (including several previous models as special cases).
- *Flexible*: level of *interaction between decoders* is a design choice.

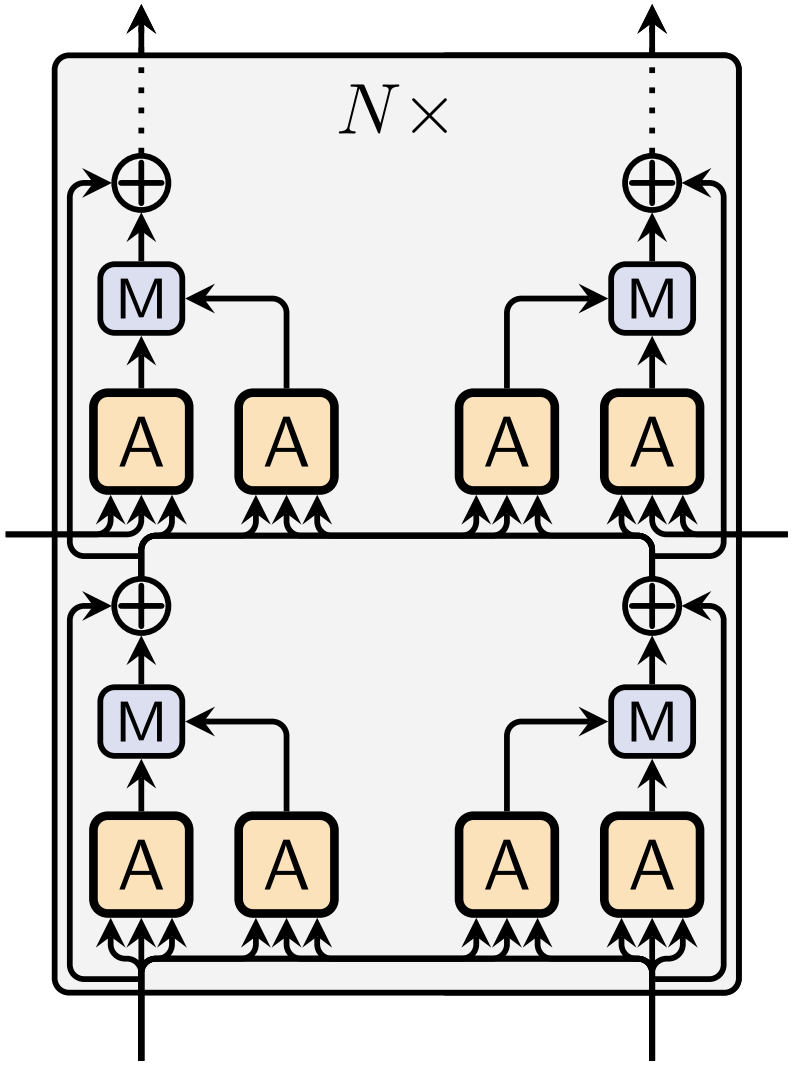




Dual-decoder Transformer



Cross dual-decoder Transformer



Parallel dual-decoder Transformer

A (Attention), **M** (Merge), **L** (LayerNorm)

Main Findings and Results

Main findings

- Dual-attention enables the decoders to effectively help each other.
- Parallel dual-attention improves both translation and transcription.
- Symmetric design is better than asymmetric one.
- Wait-*k*: Letting ASR be ahead is better than letting ST be ahead.

Results on MuST-C tst-COMMON

No	type	side	self	src	merge	epochs	de	es	fr	it	nl	pt	ro	ru	avg	WER
1	Bilingual (Inaguma et al., 2020)					50	22.91	27.96	32.69	23.75	27.43	28.01	21.90	**15.75**	25.05	12.0
2	One-to-many (Gangi et al., 2019)						17.70	20.90	26.50	18.00	20.00	22.60	-	-	-	-
3	One-to-many (Gangi et al., 2019)						16.50	18.90	24.50	16.20	17.80	20.80	15.90	9.80	17.55	-
4	independent++					25	22.82	27.20	32.11	23.34	26.67	28.98	21.37	14.34	24.60	11.6
5	par	both	✓	✓	concat	25	22.74	27.59	32.86	23.50	26.97	29.51	21.94	14.88	25.00	11.6
6	par^{R3}	both	-	✓	sum	25	22.84	27.92	32.12	23.61	27.29	29.48	21.16	14.50	24.87	11.6
7	par++	both	-	✓	sum	25	**23.63**	**28.12**	**33.45**	**24.18**	**27.55**	**29.95**	**22.87**	15.21	**25.62**	**11.4**