DIRT: Distributed Internal Regression Transformer

This research is financed by the CALCULUS project - Commonsense and Anticipation enriched Learning of Continuous representations

European Research Council Advanced Grant H2020-ERC-2017-ADG 788506

Nathan Cornille
Damien Sileo
Marie-Francine Moens

Department of Computer Science

November 9-12, 2020

Context

March 2020

Abstract: paper idea

July 2020

Negative results: paper stopped

November 2020

NAISys 2020

Motivation

💬 Improve **Natural Language Understanding**

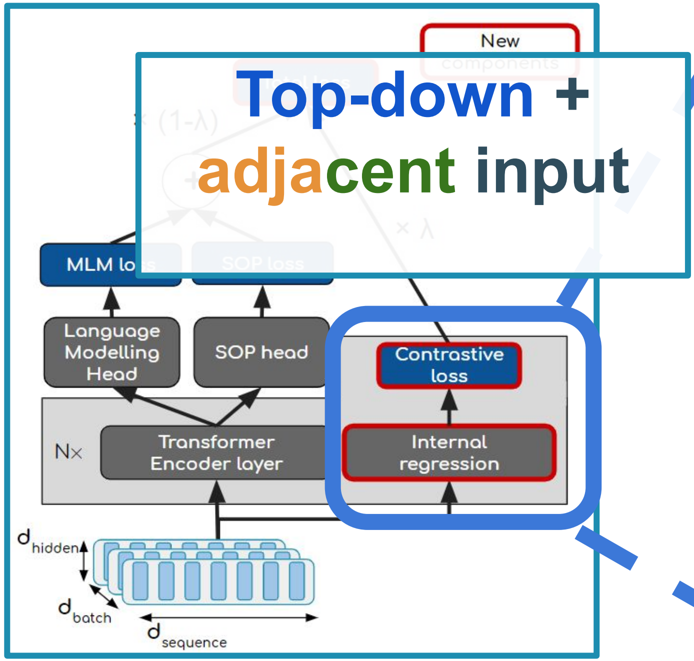
Improve quality of **general-purpose representations**

🧠 **Train with internal self-prediction loss**[1,2,3,4]

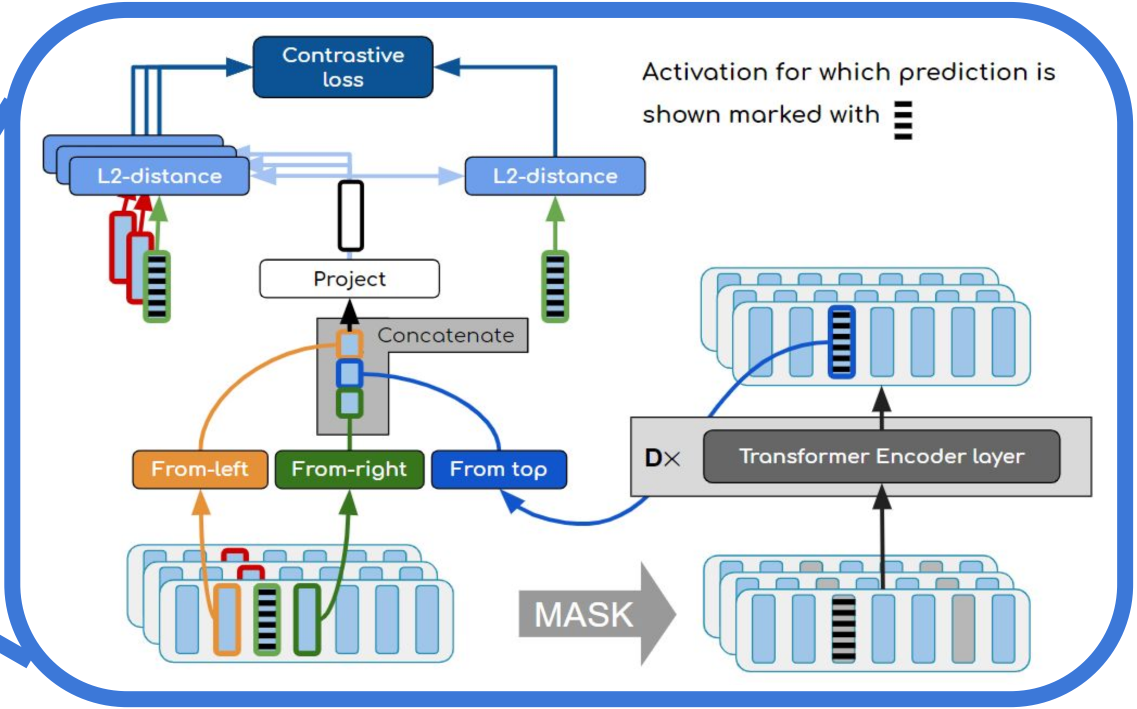
Faculty of Engineering Science,
Department of Computer Science,
HCI unit
KU LEUVEN

Implementation

Faculty of Engineering Science,
Department of Computer Science,
HCI unit



Per layer loss



Contrastive loss

- Prevent “cheating” by the model
- **Bonus**: induce desirable **“slow features”**

Results

SuperGLUE[5] benchmark

$\lambda = 0$ (baseline)	61.2±0.7
$\lambda = 0.4$	**61.5±0.7**
$\lambda = 0.9$	60.9±0.9

Negative result: no significant improvement over baseline

	Avg	BoolQ Acc.	CB Acc./F1 Avg	COPA Acc.	MultiRC $F1_\alpha$/EM Avg	ReCoRD F1/EM Avg	RTE Acc.	WiC Acc.	WSC Acc.	AX_b MCC	AX_g Acc./GPS
$\lambda = 0$ (baseline)	61.2±0.7	75.4±0.7	**71±4.5**	55±2.8	43.8±0.8	45.4±2.9	**71.7±0.9**	68.6±1	59.1±2.2	19.1±2	50.6±1.3/96.8±1.5
$\lambda = 0.4$	**61.5±0.7**	74.9±0.8	70.3±1.8	54.7±4.6	**44.4±0.3**	**48.2±1.3**	71.6±1.5	**68.7±0.9**	59.3±0.6	**19.2±1.3**	50.6±0.5/97.2±0.6
$\lambda = 0.9$	60.9±0.9	**75.7±0.7**	70.9±8.2	**55.3±2.5**	43.6±0.6	43.1±2.1	71.6±0.8	67.8±0.8	59.3±3.9	17.2±1.3	51.4±2.3/97.6±1.2
$\lambda = 1$	42±2.4	62.2±0	36.1±0	53.5±7.8	9.4±11.8	13.8±0.4	47.1±0.3	50±0	**63.5±0**	0±0	**51.7±X/100±X**
Most Frequent	47.7	62.2	36.1	55.0	30.4	32.0	52.7	50.0	**63.5**	0.0	50/100
CBoW	47.7	62.4	60.5	**63.0**	10.3	14.1	54.2	55.3	61.5	-0.4	50/100

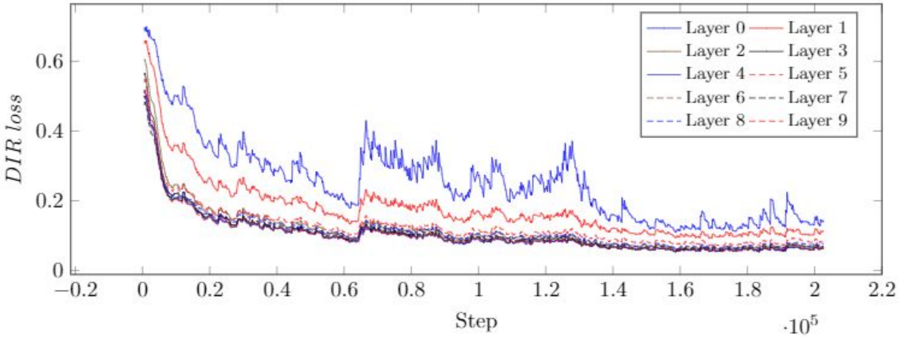
Faculty of Engineering Science,
Department of Computer Science,
HCI unit

KU LEUVEN

Internal loss

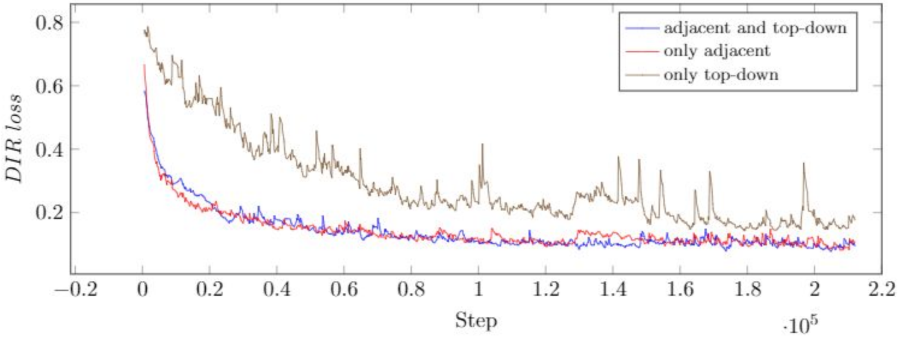
First layers are hardest to self-predict

Figure 3: Evolution of DIR loss at different layers, shown for $\lambda = 0.4$.



Top-down signal doesn't add value

Figure 4: Inner self-prediction loss for different ablations of input for self-prediction.



Faculty of Engineering Science,
Department of Computer Science,
HCI unit

KU LEUVEN

Lessons learned

Faculty of Engineering Science,
Department of Computer Science,
HCI unit

Lessons learned

- **Stepping back**
  - Actually complementary?

- **Contrastive loss red herring**
  - slow features ⇔ local input
  - **minmax objective** as alternative cheating-prevention
    - more biologically plausible too?

Faculty of Engineering Science,
Department of Computer Science,
HCI unit

References

1. K. L. Downing. Predictive models in the brain.Connection Science, 21(1):39–74, 3 2009.
2. L. Grisoni, B. Mohr, and F. Pulvermüller. Prediction mechanisms in motor and auditory areas and their role in sound perception and language understanding.NeuroImage, 199:206–216, 10 2019.
3. A. Modi, I. Titov, V. Demberg, A. Sayeed, and M. Pinkal. Modeling Semantic Expectation: Using Script Knowledge for Referent Prediction.Transactions of the Association for Computational Linguistics, 5:31–44, 12 2017.
4. A. F. Morse, H. Svensson, and T. Ziemke. Representation as Internal Simulation : A Minimalistic Robotic Model, 2009.
5. A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, andS. R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. 5 2019.

DIRT: Distributed Internal Regression Transformer

- Context
- Motivation
- Implementation
- Results
- Lessons learned

This research is financed by the CALCULUS project - Commonsense and Anticipation enriched Learning of Continuous representations

European Research Council Advanced Grant H2020-ERC-2017-ADG 788506

Nathan Cornille
Damien Sileo
Marie-Francine Moens

Department of Computer Science

November 9-12, 2020

Goal hierarchy

- 🌍 Better world
- 📄 **Internal self-prediction loss (this work)**

Faculty of Engineering Science,
Department of Computer Science,
HCI unit

Goal hierarchy

- 🌍 Better world
- ⏱️ Increased automation
- 💬 Better language-understanding machines
- Better general-purpose NLU representations
- 🧠 neuro-for-AI
- 📄 **Internal self-prediction loss (this work)**

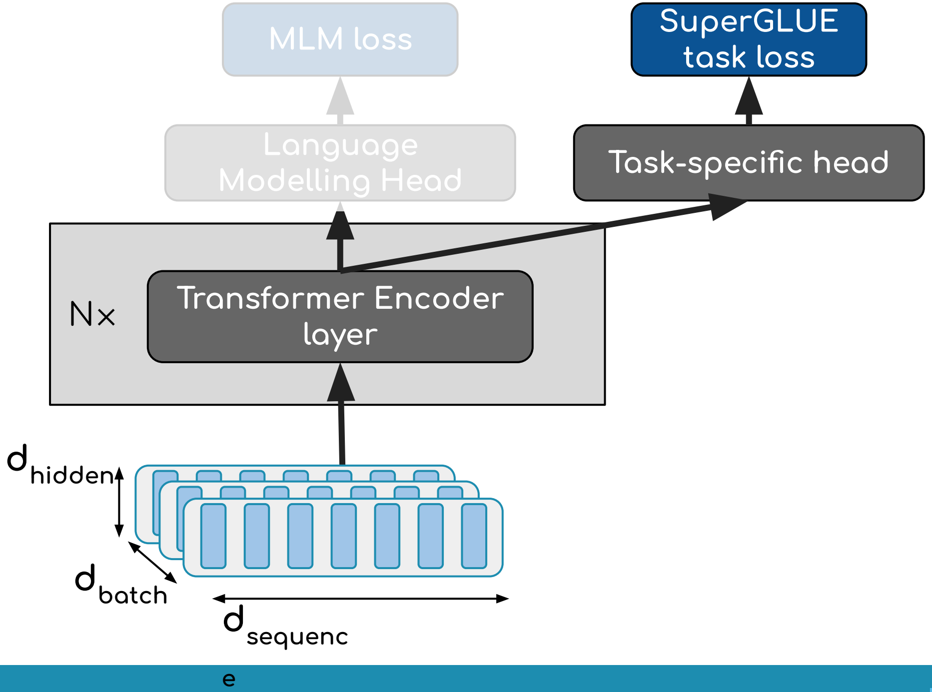
Faculty of Engineering Science,
Department of Computer Science,
HCI unit

KU LEUVEN

Baseline Transformer

Step 1: Pretraining on massive unlabeled data

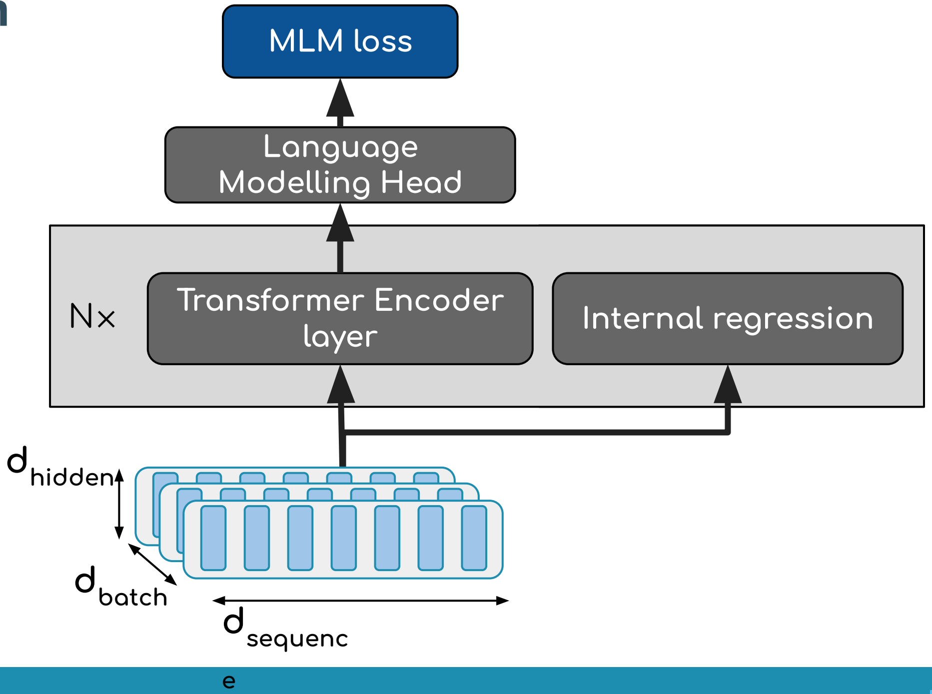
Step 2: Finetune pretrained model on variety of downstream tasks



Faculty of Engineering Science,
Department of Computer Science,
HCI unit
KU LEUVEN

Proposed extension

Distributed Internal Regression Transformer (DIRT)



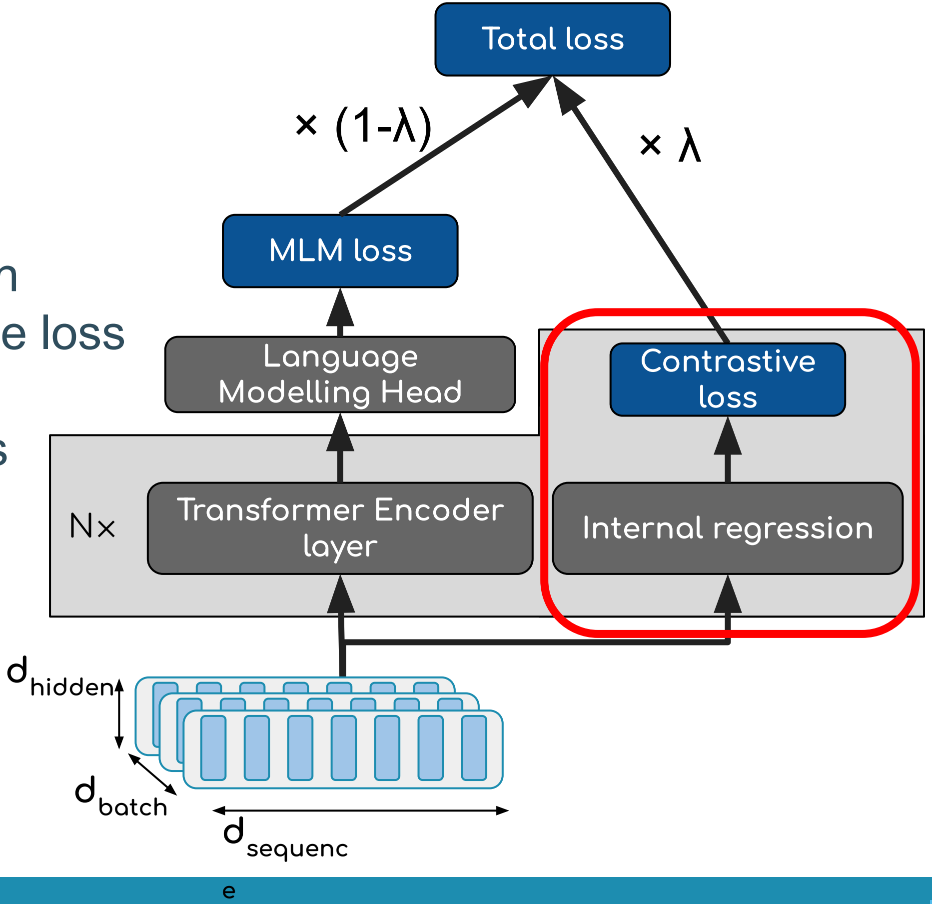
Faculty of Engineering Science,
Department of Computer Science,
HCI unit
KU LEUVEN

DIRT-as-objective

1. Start with pretrained weights

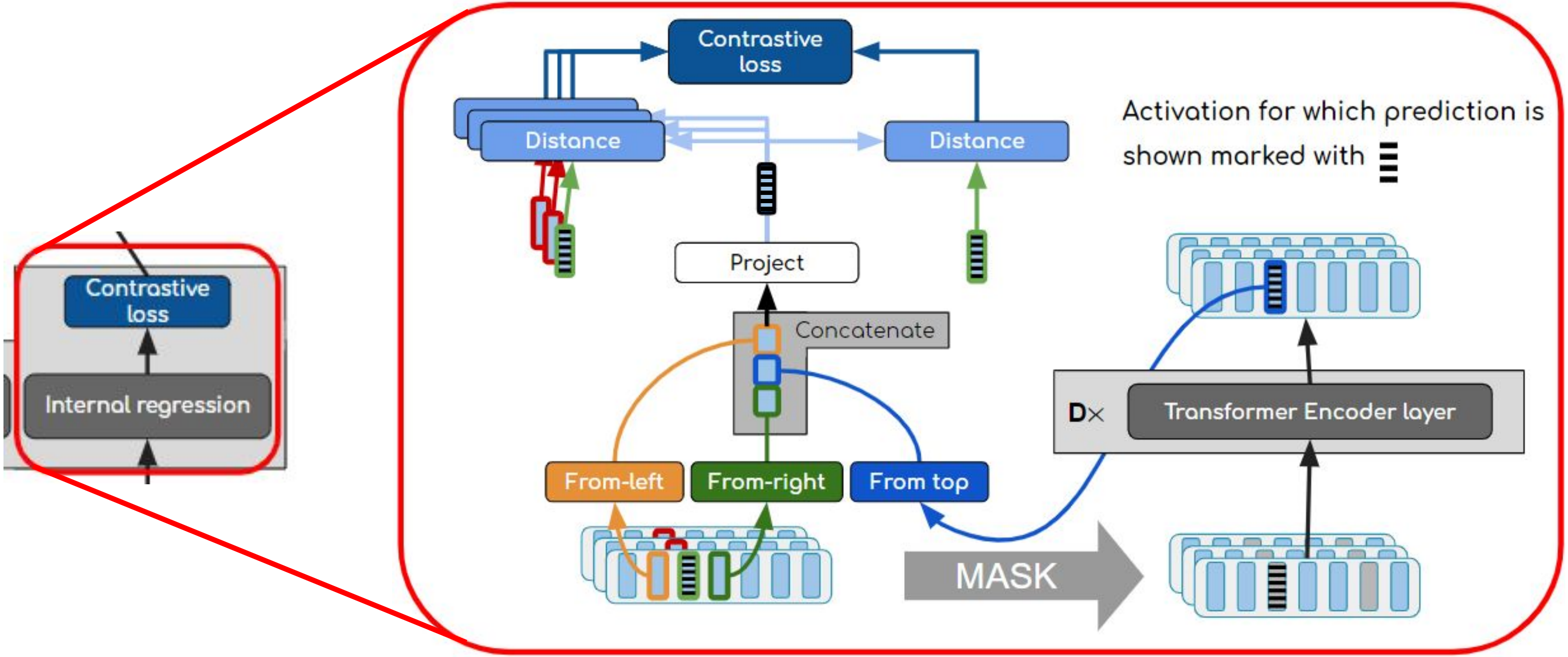
2. Do ***additional pretraining*** with anticipation-inspired contrastive loss

3. Finetune on downstream tasks



Faculty of Engineering Science,
Department of Computer Science,
HCI unit

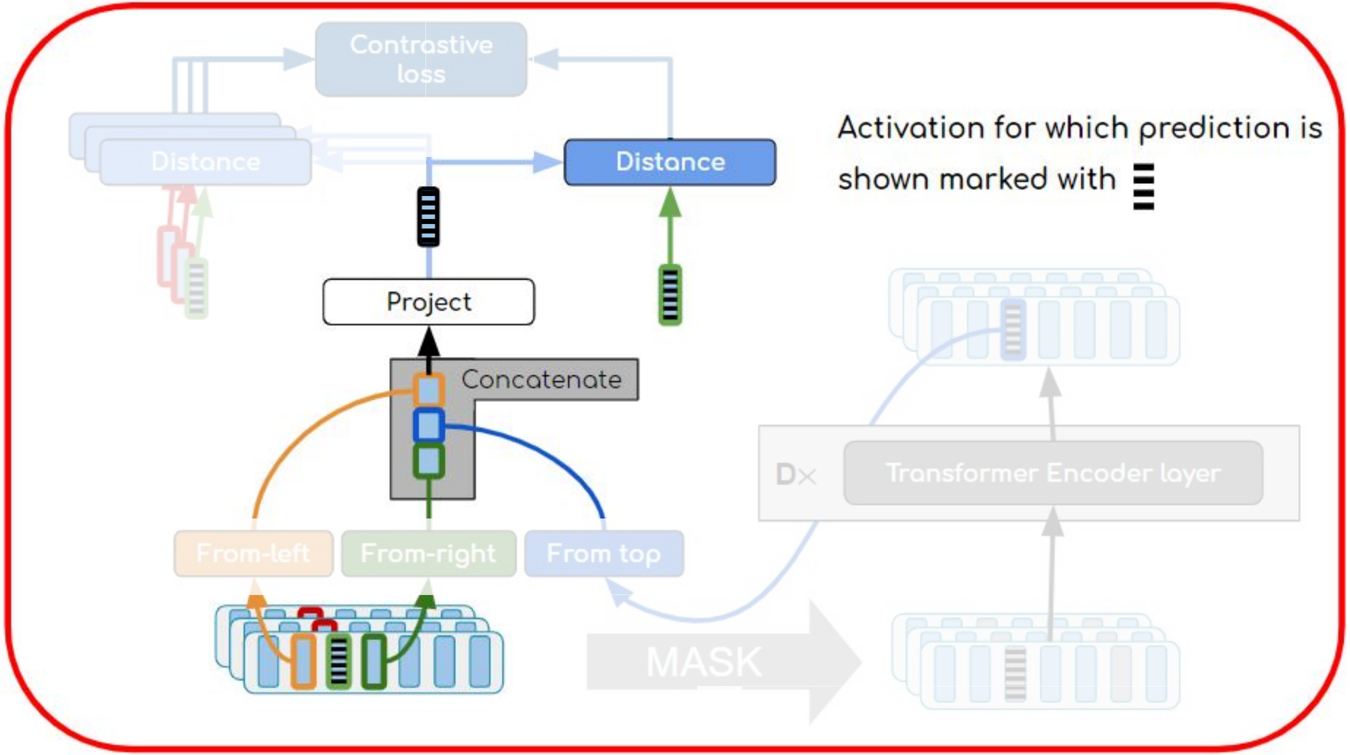
Detailed view



Detailed view

Porting to NLU model

Autoregressing to masked internal states



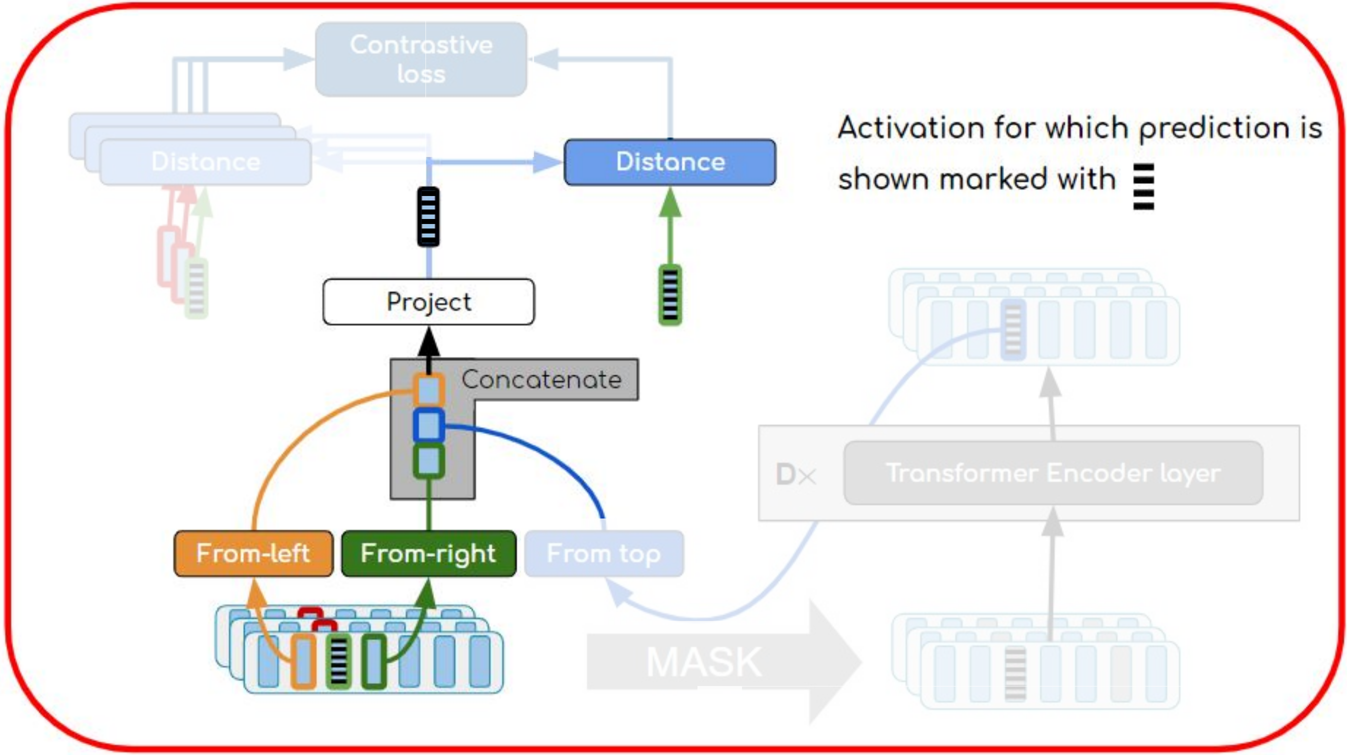
Faculty of Engineering Science,
Department of Computer Science,
HCI unit
KU LEUVEN

Detailed view

Porting to NLU model

Autoregressing to masked internal states

Input from neighbouring timesteps + top-down timesteps



Faculty of Engineering Science,
Department of Computer Science,
HCI unit

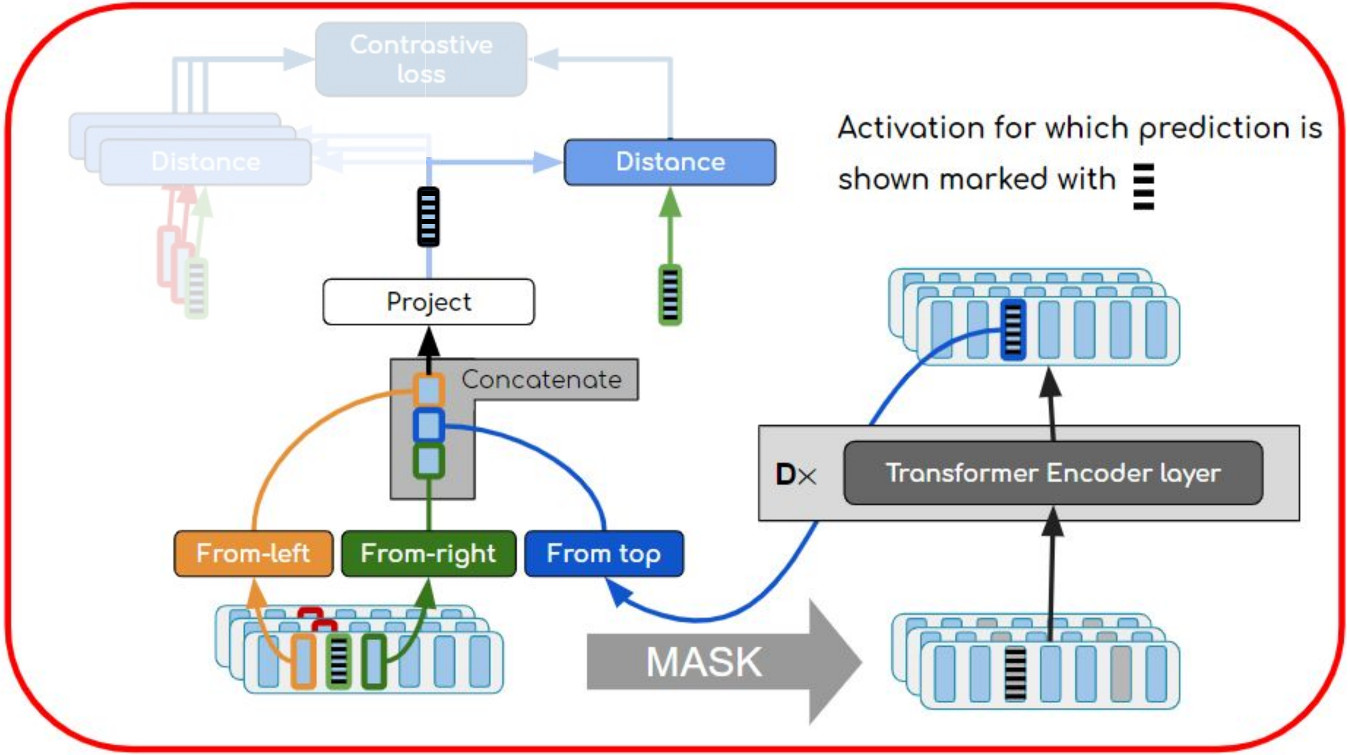
KU LEUVEN

Detailed view

Porting to NLU model

Autoregressing to masked internal states

Input from neighbouring timesteps + top-down timesteps



Faculty of Engineering Science,
Department of Computer Science,
HCI unit
KU LEUVEN

Detailed view

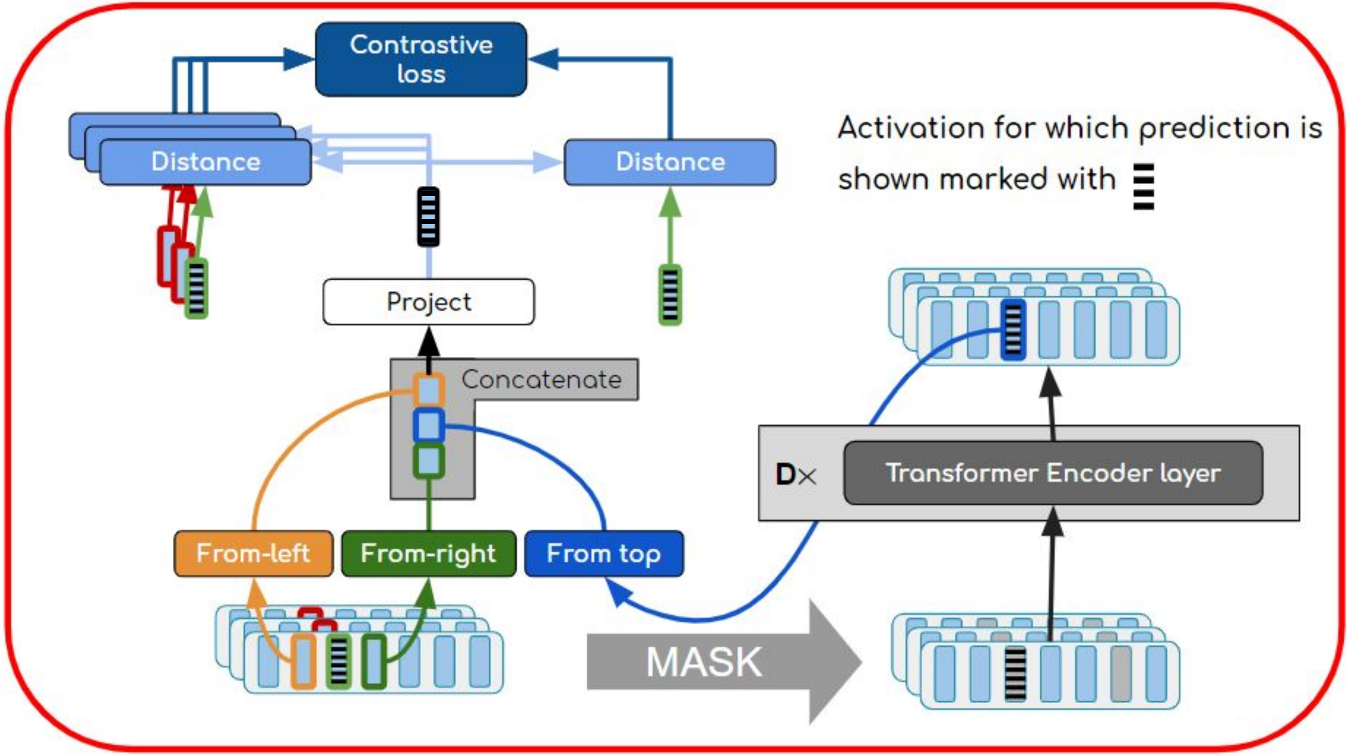
Porting to NLU model

Internal losses

- Grounding via contrasting

Autoregressing to masked internal states

Input from neighbouring timesteps + top-down timesteps



Faculty of Engineering Science,
Department of Computer Science,
HCI unit

KU LEUVEN