Why Transformers Work.

I'm Vincent
Ask Me Anything [tm]
Think about all the actions in this dialogue.

😊 Hello

😊 Hi

😊 I’d like a pizza!

😊 What kind?

😊 By the way, are you a human?

😊 No, I’m a bot.

😊 But what kind of pizza?
intent 😊 Hello

entity

intent 😊 I'd like a pizza!

What kind?

intent 😊 By the way,

are you a human?

No, I'm a bot.

But what kind of pizza?

*More info blablabla*
intent 😊 Hello

intent 😊 I'd like a pizza!

intent 😊 By the way, are you a human?

intent 😊 No, I'm a bot. But what kind of pizza?
I wanna buy a pizza.
I wanna buy a pizza.
1. as far as $x_i$ is concerned
1. as far as $x_i$ is concerned
2. this is what we need to pay attention to

*More info blablabla*
1. as far as $x_i$ is concerned

2. this is what we need to pay attention to

3. this allows us to re-weigh and this operation hopefully gives us extra context

*More info blablabla*
Bank of the river.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

Money on the bank.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

*More info blablabla*
Bank of the river.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

If \( v_1 \) and \( v_2 \) share info

\[ v_1 \cdot v_2 \text{ is big} \]

Money on the bank.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

If \( v_1 \) and \( v_2 \) don't

\[ v_1 \cdot v_2 \approx 0 \]

*More info blablabla*
Bank of the river.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

Money on the bank.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]
Bank of the river:
\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]
\[ V_1 \quad v_2 \quad v_3 \quad v_4 \]

Money on the bank:
\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]
\[ V_1 \quad v_2 \quad v_3 \quad v_4 \]

NORMALISE

*More info blablabla*
Bank of the river:

\[ t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \]
\[ v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \]

\[ v_1 \cdot v_1 \rightarrow v_1 \cdot v_2 \rightarrow v_1 \cdot v_3 \rightarrow v_1 \cdot v_4 \]
\[ w_{11} \rightarrow w_{12} \rightarrow w_{13} \rightarrow w_{14} \]

Money on the bank:

\[ t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \]
\[ v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \]

\[ v_4 \cdot v_1 \rightarrow v_4 \cdot v_2 \rightarrow v_4 \cdot v_3 \rightarrow v_4 \cdot v_4 \]
\[ w_{41} \rightarrow w_{42} \rightarrow w_{43} \rightarrow w_{44} \]

\[ v_1^* = \sum_j w_{1j} v_j \]
\[ v_4^* = \sum_j w_{4j} v_j \]
Bank of the river.

\[ t_1 \quad t_2 \quad t_3 \quad t_4 \]

\[ v_1 \quad v_2 \quad v_3 \quad v_4 \]

\[ v^*_i = \sum_j w_{ij} v_j \]
Bank of the river.

\[ t_1, t_2, t_3, t_4 \]

\[ v_1, v_2, v_3, v_4 \]

\[ v_1, v_2, v_3, v_4 \]

\[ w_{11}, w_{12}, w_{13}, w_{14} \]

\[ v_1^* = \sum_j w_{1j} v_j \]

Reweighing based on time-distance

\[ \sigma = 0.1 \]
Bank of the river:

\[ t_1, t_2, t_3, t_4 \]

\[ v_1, v_2, v_3, v_4 \]

\[ w_{11}, w_{12}, w_{13}, w_{14} \]

\[ v_i^* = \sum_j w_{ij} v_j \]

Reweighing based on embedding similarity

Reweighing based on time-distance

*More info blablabla*
Bank of the river.

\[ t_1, t_2, t_3, t_4 \]
\[ v_1, v_2, v_3, v_4 \]

\[ v_1 v_1, v_1 v_2, v_1 v_3, v_1 v_4 \]
\[ w_1, w_2, w_3, w_4 \]

\[ v_1^* = \sum_j w_j v_j \]

Money on the bank.

\[ t_1, t_2, t_3, t_4 \]
\[ v_1, v_2, v_3, v_4 \]

\[ v_4 v_1, v_4 v_2, v_4 v_3, v_4 v_4 \]
\[ w_4, w_4, w_4, w_4 \]

\[ v_4^* = \sum_j w_j v_j \]
Bank of the river:

\[ t_1, t_2, t_3, t_4 \]

\[ v_1, v_2, v_3, v_4 \]

Money on the bank:

\[ t_1, t_2, t_3, t_4 \]

\[ v_1, v_2, v_3, v_4 \]

\[ v_1^* = \sum_j w_{1j} v_j \]

\[ v_4^* = \sum_j w_{4j} v_j \]

Let's move this idea into a layer.
SELF ATTENTION BLOCK

V_1 \rightarrow \rightarrow V_1^*
V_2 \rightarrow \rightarrow V_2^*
V_3 \rightarrow \rightarrow V_3^*
V_4 \rightarrow \rightarrow V_4^*
V_5 \rightarrow \rightarrow V_5^*

*More info blablabla
SELF ATTENTION BLOCK

\[
\begin{bmatrix}
  v_1 \\
  \vdots \\
  v_5
\end{bmatrix}
\xrightarrow{\text{MATMUL}}
\begin{bmatrix}
  v_1^* \\
  v_5^*
\end{bmatrix}
\xrightarrow{\text{NORMALISE}}
\begin{bmatrix}
  v_1^* \\
  v_2^* \\
  v_3^* \\
  v_4^* \\
  v_5^*
\end{bmatrix}
\xrightarrow{\text{MATMUL}}
\begin{bmatrix}
  v_1 \\
  \vdots \\
  v_5
\end{bmatrix}
\]
More info blablabla
It's a more elaborate way to do stuff like this.
Bank of the river.

\[ t_1, t_2, t_3, t_4 \]

\[ V_1, V_2, V_3, V_4 \]

\[ V_{11}, V_{12}, V_{13}, V_{14} \]

\[ v_1^* = \sum_j w_{ij} v_j \]

\[ \begin{bmatrix} v_1^* \\ \vdots \\ v_k^* \end{bmatrix} \]

\[ \uparrow \]

SELF ATTENTION BLOCK

\[ \begin{bmatrix} v_1 \\ \vdots \\ v_k \end{bmatrix} \]

\[ \uparrow \]
Bank of the river:

\[v_1 = \sum_j w_{ij} v_j\]
Bank of the river:

\[ t_1, t_2, t_3, t_4 \]

\[ v_1, v_2, v_3, v_4 \]

\[ v_1^*, v_2^*, v_3^*, v_4^* \]

\[ w_{11}, w_{12}, w_{13}, w_{14} \]

\[ v^*_1 = \sum_j w_{ij} v_j \]

TRANSFORMER LAYER:

\[ \begin{bmatrix} v_1^* \\ \vdots \\ v_n^* \end{bmatrix} \]

EXTRA LAYERS

SELF ATTENTION BLOCK

EXTRA LAYERS

\[ \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \]
*Glancing over some details
the goal here is intuition
in the interest of time*
Needs a lot of data to stop looking at its previous neighbor.

Starts out by looking at similarity of pretrained embeddings.
Needs a lot of data to stop looking at its previous neighbor.

Hard to parallelise.

Starts out by looking at similarity of pretrained embeddings.

More parallel options.

RNN

*More info blablabla
intent 😊 Hello

intent 😊 I'd like a pizza!

intent 😊 By the way, are you a human?

intent 😊 No, I'm a bot.

intent 😊 But what kind of pizza?
Q: So how does Rasa find these intents and entities?

A: DIET

Dual Intent & Entity Transformer
play

ping

pong

CLS
Q: So how does Rasa determine what actions to take?

A: TED
Transformer Embedding Dialogue
TED policy

"I'd like a pizza."
TED policy

"I'd like a pizza."

order

intent

[0]

[0]
TED policy

"I'd like a pizza."
TED policy

"I'd like a pizza."

intent
- [0 0]
- [0 1]
- [0 0]

entity
- [0 0]
- [1 0]
- [0 0]

slot
- [0 0]
- [0 0]

order
pizza
adress
"I'd like a pizza."

TED policy

intent

- [0 1 0]
- [1 0 0]

entity

- [0 1 0]

order

- [0 1 0]

pizza

slot

- [0 1 0]

adress

prev action

- [1 0 0]

listen

$at-1$
"I'd like a pizza."

TED policy

```
intent entity
[0 0 1 0] [0 0 1]
order pizza
slot prev action
[0 0 1] [1 0 0]
address listen
```

$f_t$
TED policy

"I'd like a pizza." → intent
   ↘
order
   ↘
slot
   ↘
address
   ↘
pizza
   ↘
prev
action
   ↘
listen
   ↘
model

*More info blablbla
TED policy

\[ f_t \rightarrow MODEL \rightarrow a_t \]

\[ f_{t+1} \rightarrow a_{t+1} \]

\[ f_{t+2} \rightarrow a_{t+2} \]

*More info blablabla*
More info blablabla
UNIDIRECTIONAL TRANSFORMER

\[ f_0 \rightarrow \hat{f}_0 \rightarrow \text{DENSE} \rightarrow \text{SIM} \rightarrow \text{DENSE} \rightarrow a_0 \]

\[ f_1 \rightarrow \hat{f}_1 \rightarrow \text{DENSE} \rightarrow \text{SIM} \rightarrow \text{DENSE} \rightarrow a_1 \]

\[ f_2 \rightarrow \hat{f}_2 \rightarrow \text{DENSE} \rightarrow \text{SIM} \rightarrow \text{DENSE} \rightarrow a_2 \]

\[ f_3 \rightarrow \hat{f}_3 \rightarrow \text{DENSE} \rightarrow \text{SIM} \rightarrow \text{DENSE} \rightarrow a_3 \]

only look back
attention (future) = 0

LOSS

*More info blablabl*
Demo time
config.yml

policies:
- name: MemoizationPolicy
- name: TEDPolicy
  epochs: 200
  max_history: 1
- name: MappingPolicy

Your input -> start counting
  Countdown! ETA 10 🎵
Your input -> ok
  Countdown! ETA 5 🎵
Your input -> ok
  Countdown! ETA 5 🎵
Your input -> ok
  Countdown! ETA 5 🎵
Your input -> ok
  Countdown! ETA 5 🎵

*More info blablabla*
config.yml

policies:
  - name: MemoizationPolicy
  - name: TEDPolicy
    epochs: 200
    max_history: 3
  - name: MappingPolicy

Your input -> count
Countdown! ETA 10 🎶
Your input -> ok
Countdown! ETA 9 🎶
Your input -> ok
Countdown! ETA 8 🎶
Your input -> ok
Countdown! ETA 7 🎶
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
Countdown! ETA 6 🎶
Your input -> ok
Countdown! ETA 5 🎶
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
Your input ->
Countdown! ETA 10 🎶
Your input -> count
  Countdown! ETA 10 🎵
Your input -> ok
  Countdown! ETA 9 🎵
Your input -> ok
  Countdown! ETA 8 🎵
Your input -> are you a bot?
  I am a bot, not a human, powered by Rasa.
  Countdown! ETA 7 🎵
Your input -> ok
  Countdown! ETA 6 🎵
Your input -> are you a bot?
  I am a bot, not a human, powered by Rasa.
  Countdown! ETA 5 🎵
Your input -> ok
  Countdown! ETA 4 🎵
Your input -> are you a bot?
  I am a bot, not a human, powered by Rasa.
  Countdown! ETA 3 🎵
Your input -> ok
  Countdown! ETA 2 🎵
Your input -> are you a bot?
  I am a bot, not a human, powered by Rasa.
  Countdown! ETA 1 🎵
Your input -> ok
  End of countdown!

config.yml

policies:
  - name: MemoizationPolicy
  - name: TEDPolicy
    epochs: 200
    max_history: 10
  - name: MappingPolicy

*More info blablabla
config.yml

Using LSTM policies:
- name: MemoizationPolicy
- name: KerasPolicy
  epochs: 200
- name: MappingPolicy

Your input -> count
Countdown! ETA 10 🎵
Your input -> ok
Countdown! ETA 9 🎵
Your input -> ok
Countdown! ETA 8 🎵
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
Countdown! ETA 7 🎵
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
Countdown! ETA 6 🎵
Your input -> are you a bot?
Countdown! ETA 5 🎵
Your input -> ok
Countdown! ETA 4 🎵
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
Countdown! ETA 3 🎵
Your input -> are you a bot?
I am a bot, not a human, powered by Rasa.
The real problem though.

stories that we optimise for

ensure that this overlap is big
i.e. listen to users, keep labelling data!

stories that users generate
While I have you here: check the Rasa Algorithm Whiteboard!
Our algorithms are explained in more detail! I also take requests for content.
While I have you here: check out WHATLIES

```python
orig_chart = emb.plot_interactive('man', 'woman')
new_ts    = emb | (emb['king'] - emb['queen'])
new_chart = new_ts.plot_interactive('man', 'woman')
```

It’s an open source package for visualising word embeddings. Soon: features for detecting bias. Feedback is appreciated!
Get in touch!

Vincent D. Warmerdam
Research Advocate
v.warmerdam@rasa.com
Appendix
- ORDER DOES NOT MATTER
- NO TRAINABLE WEIGHTS (YET)
- DEPENDS ON PRETRAINED QUALITY
- TASK INDEPENDENT
How's about we add us some trainable weights?
More info blablabla
MULTI HEAD ATTENTION

\[ \sum_{j} q_i \cdot k_j = 1 \]

It's a more elaborate way to do stuff like this ->

*More info blablabla
Figure 1: The Transformer - model architecture.
We use this encoder bit.

Figure 1: The Transformer - model architecture.