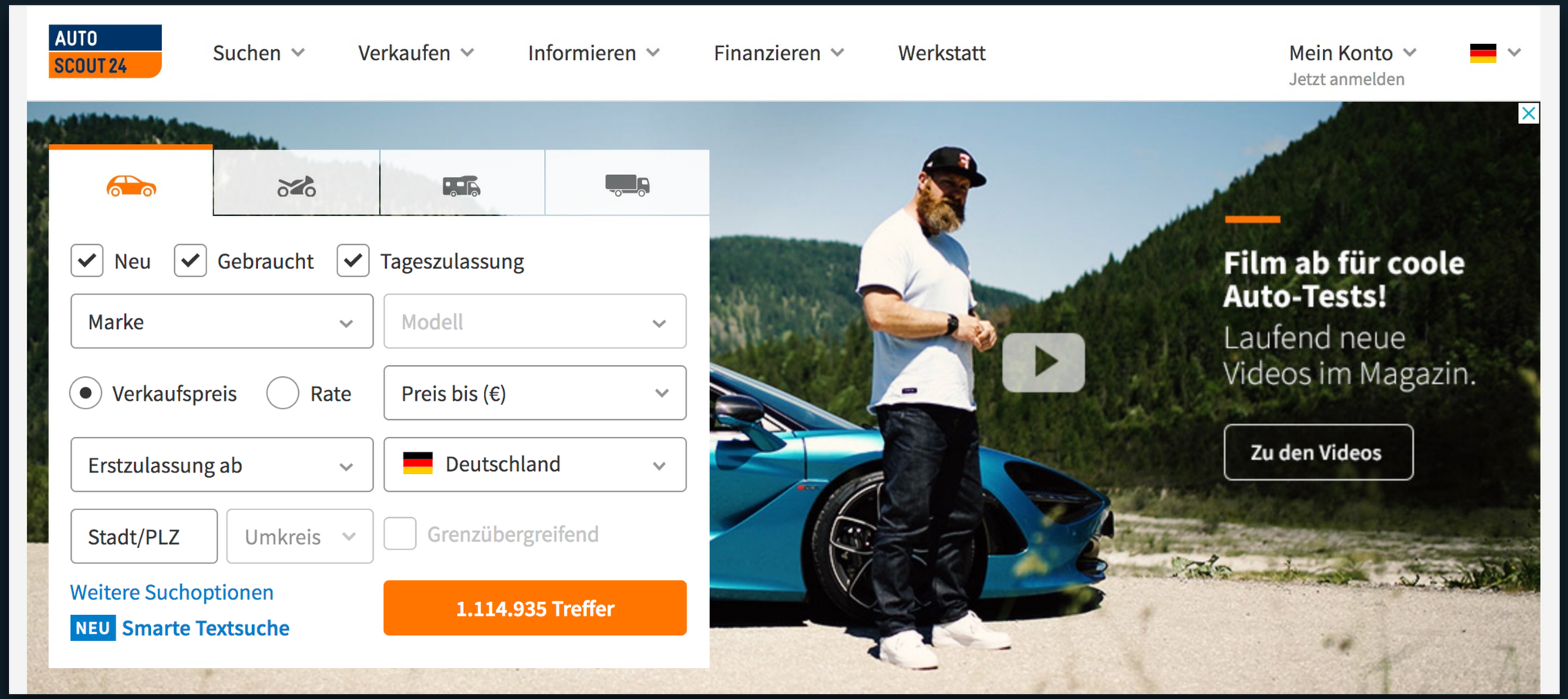
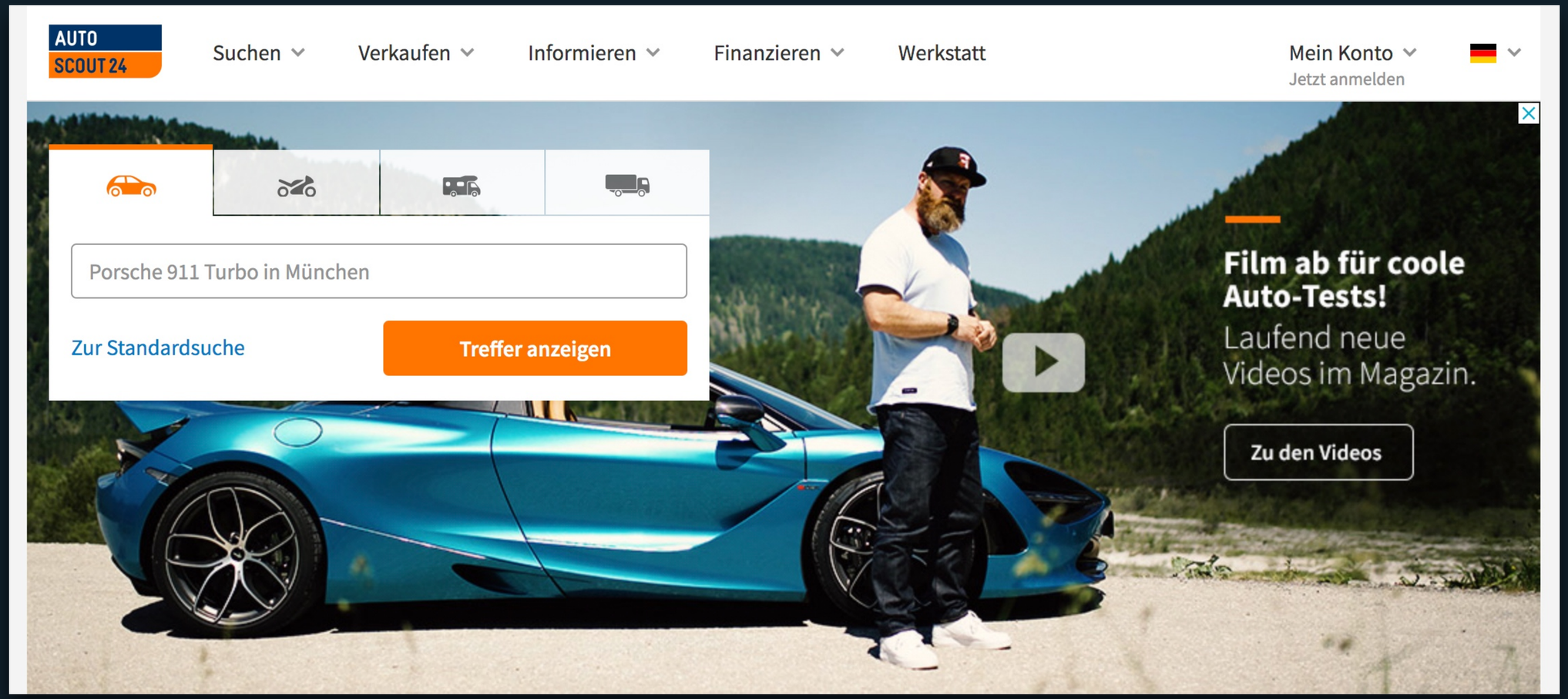
MARKUS LUDWIG | DATA COUNCIL | BARCELONA 2019

IMPROVING SEARCH WITH NATURAL LANGUAGE PROCESSING AND DEEP LEARNING

SCOUT 24


AUTO SCOUT 24
Suchen
Verkaufen
Informieren
Finanzieren
Werkstatt
Mein Konto
Jetzt anmelden
Neu
Gebraucht
Tageszulassung
Marke
Modell
Verkaufspreis
Rate
Preis bis (€)
Erstzulassung ab
Deutschland
Stadt/PLZ
Umkreis
Grenzübergreifend
Weitere Suchoptionen
NEU Smarte Textsuche
1.114.935 Treffer
Film ab für coole Auto-Tests!
Laufend neue Videos im Magazin.
Zu den Videos


AUTO SCOUT 24
Suchen
Verkaufen
Informieren
Finanzieren
Werkstatt
Mein Konto
Jetzt anmelden
Porsche 911 Turbo in München
Zur Standardsuche
Treffer anzeigen
Film ab für coole Auto-Tests!
Laufend neue Videos im Magazin.
Zu den Videos

ANATOMY OF A QUERY

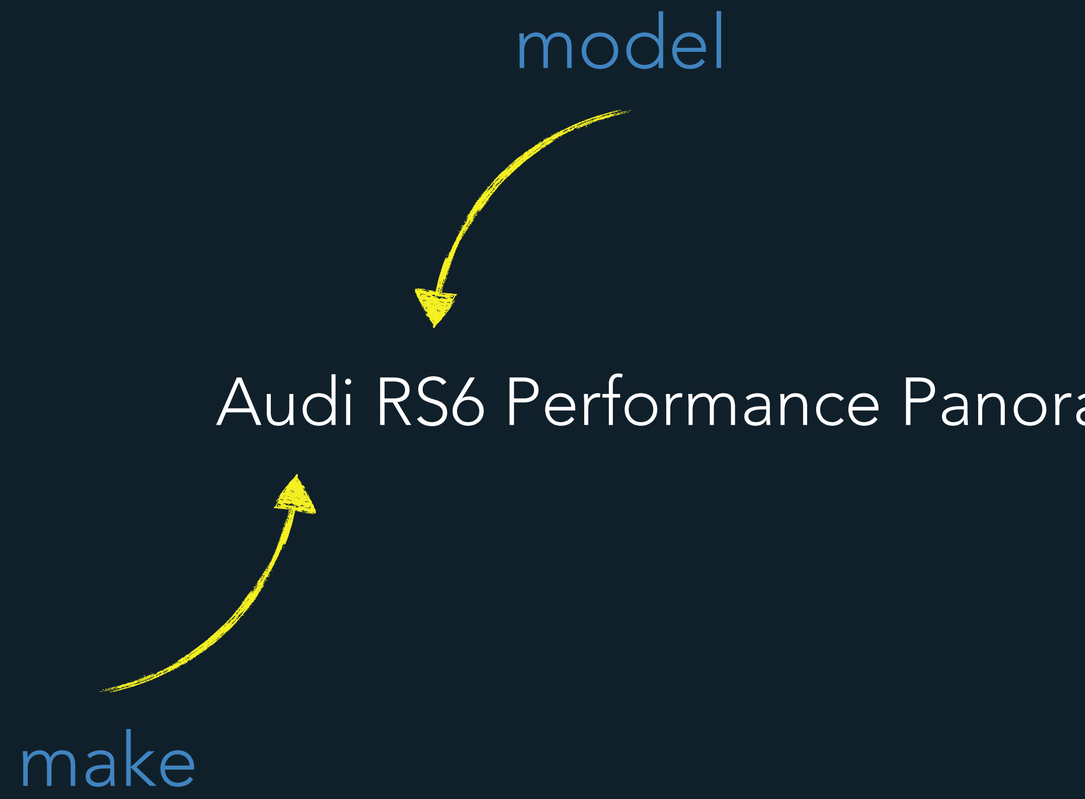
Audi RS6 Performance Panorama 2018

ANATOMY OF A QUERY

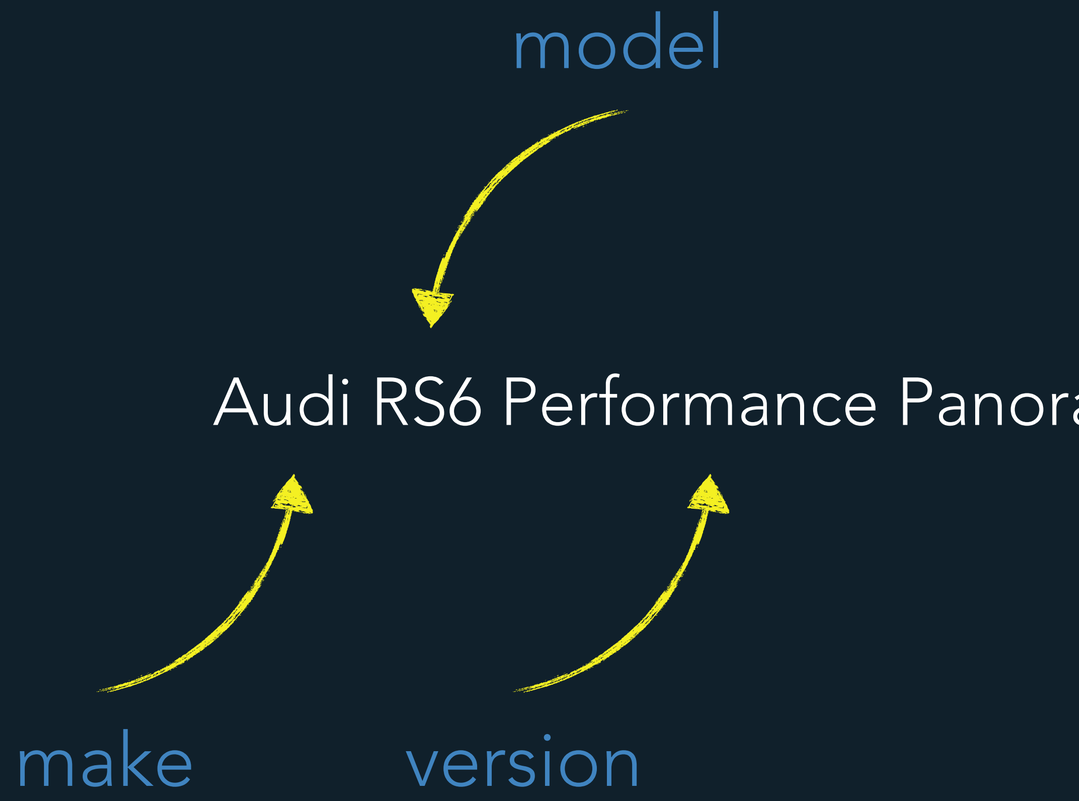
Audi RS6 Performance Panorama 2018

make

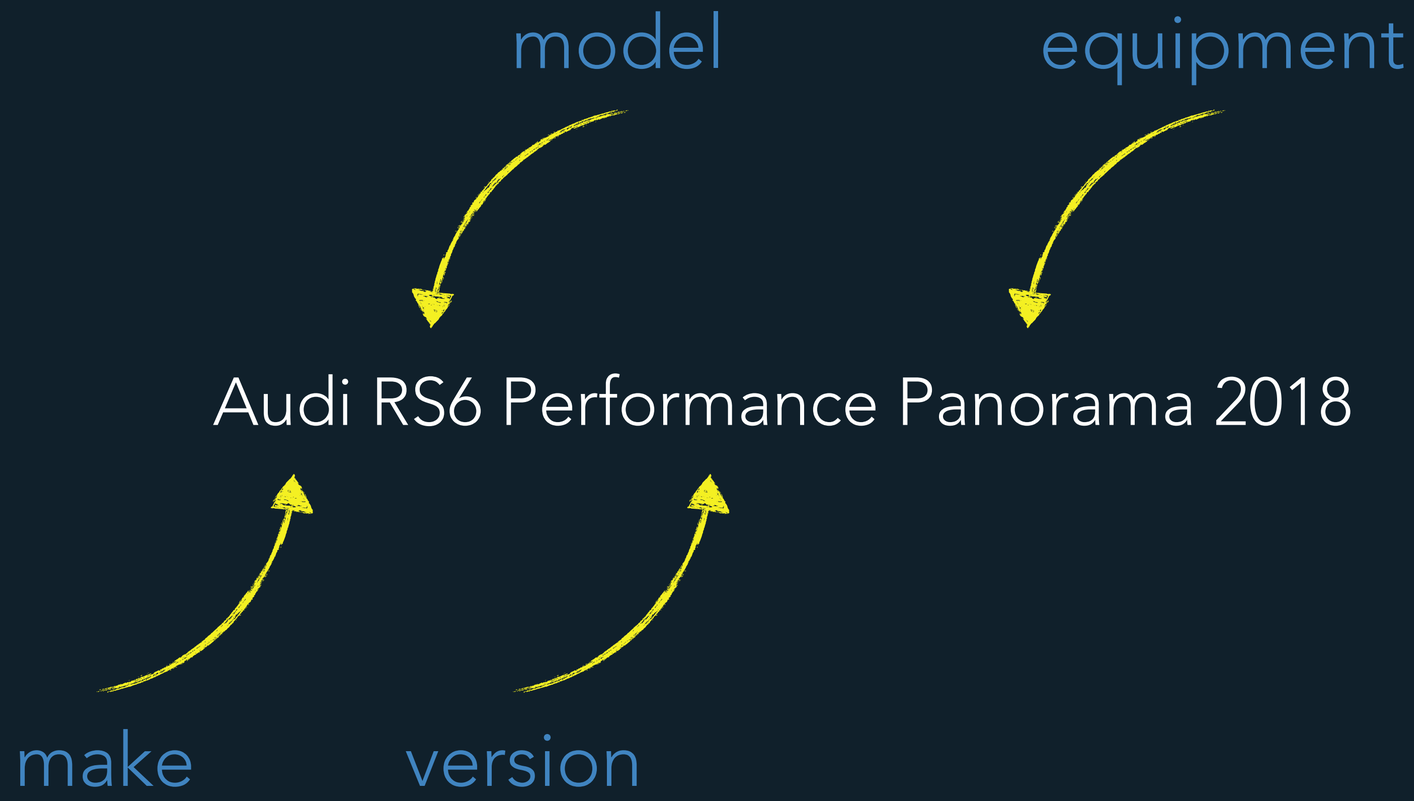
ANATOMY OF A QUERY



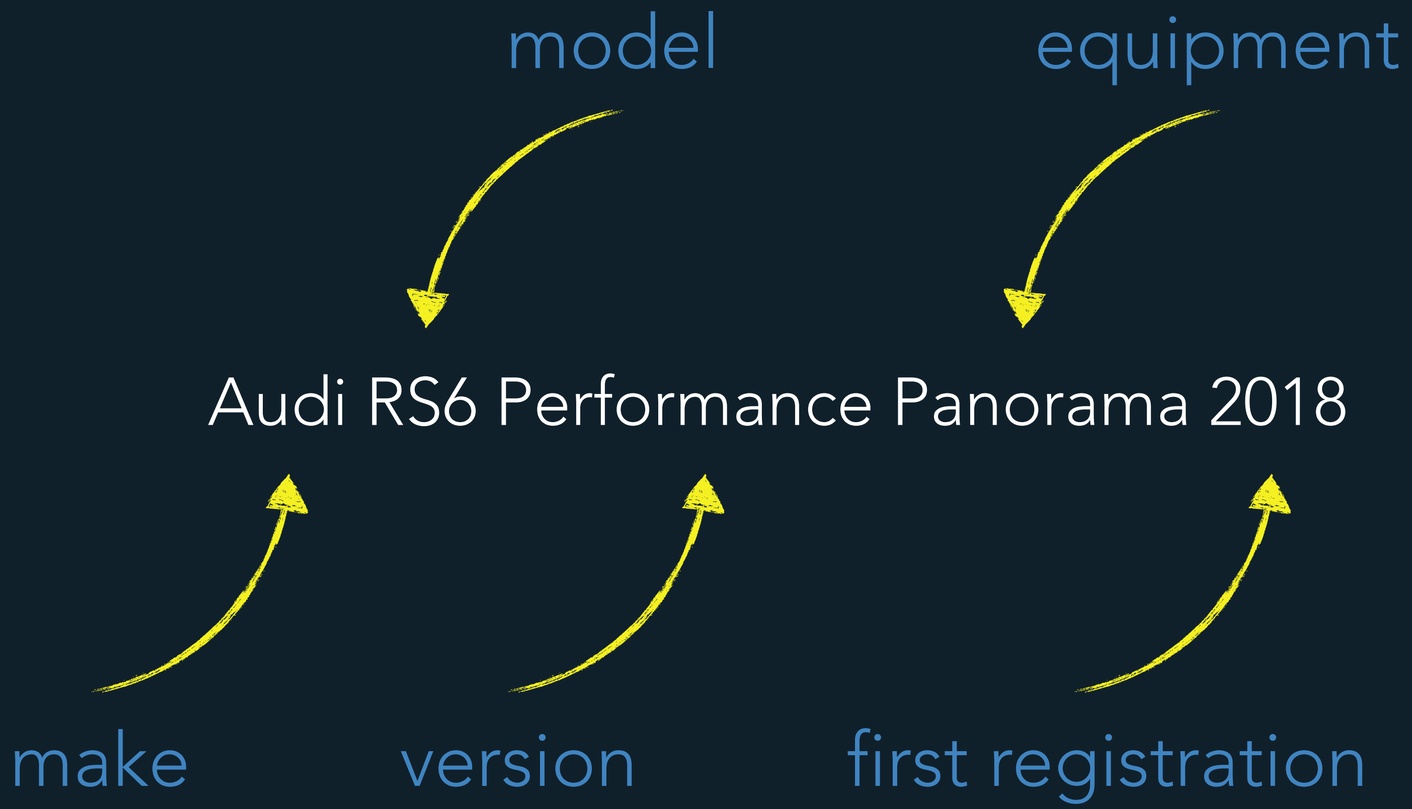
ANATOMY OF A QUERY



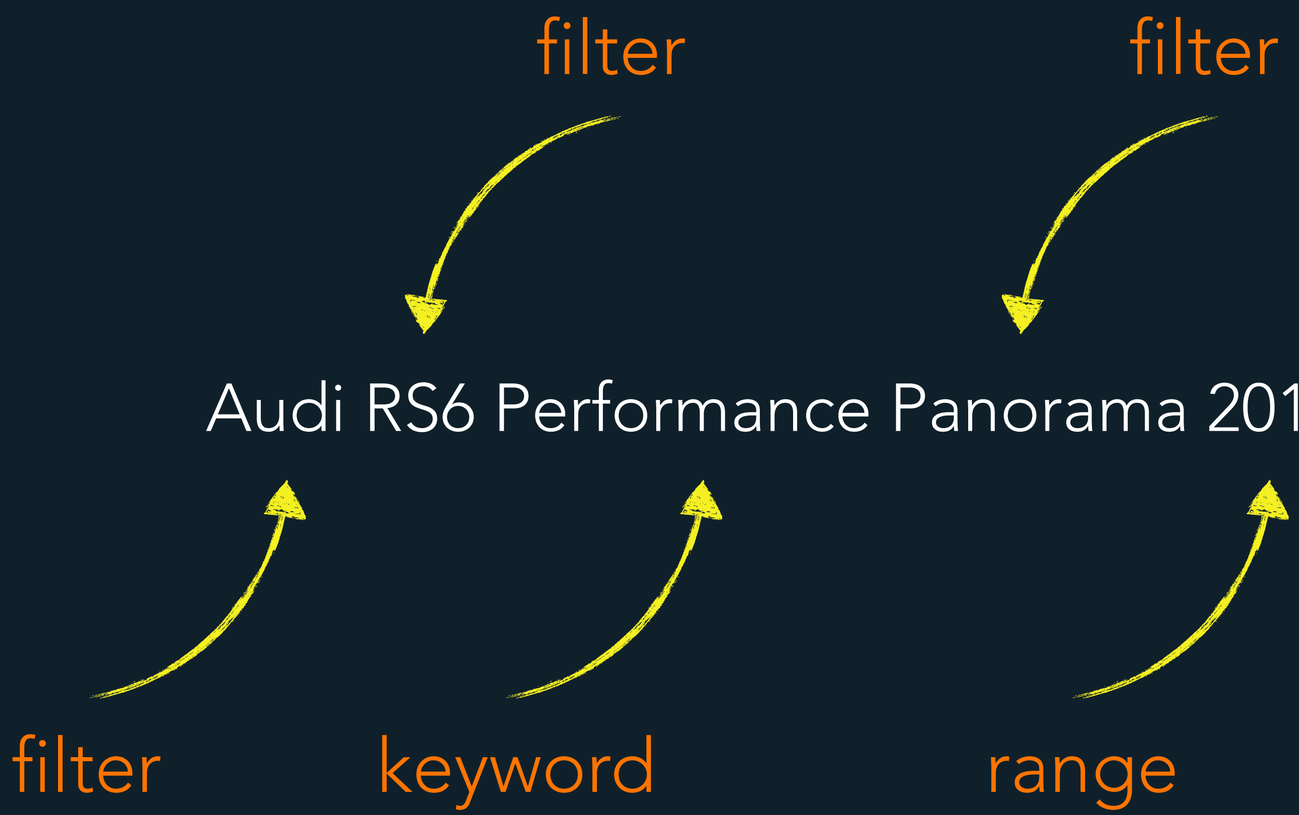
ANATOMY OF A QUERY



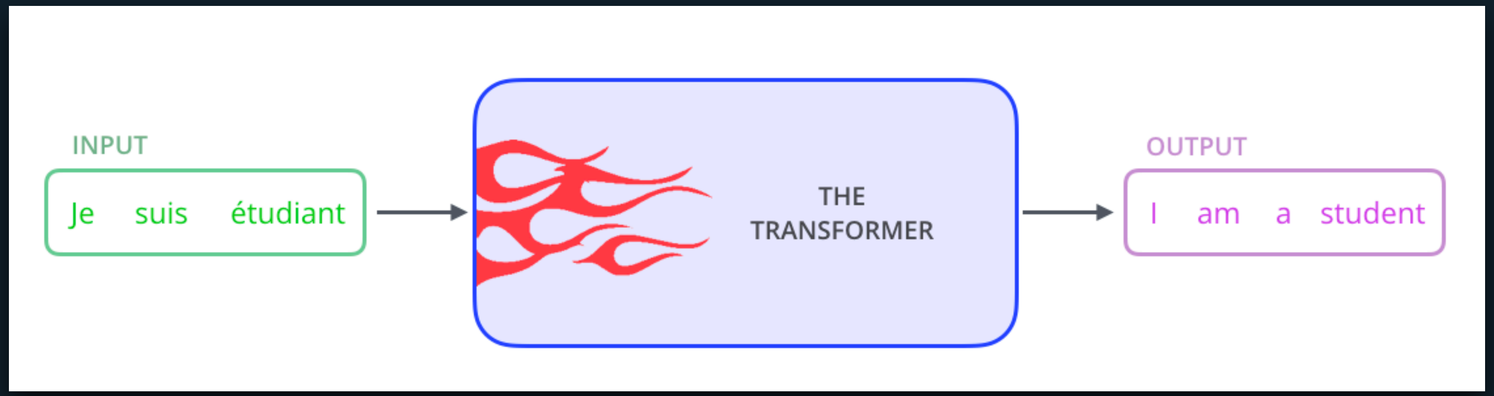
ANATOMY OF A QUERY



ANATOMY OF A QUERY



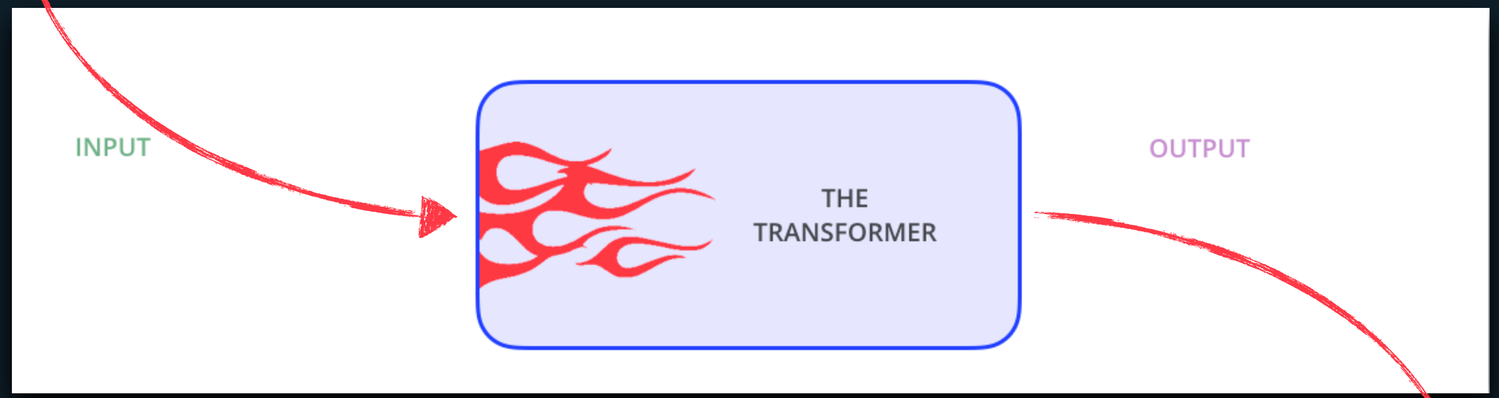
SEQUENCE-TO-SEQUENCE LEARNING



Source: http://jalammar.github.io/illustrated-transformer/

SEQUENCE-TO-SEQUENCE LEARNING

natural language



structured query

Source: http://jalammar.github.io/illustrated-transformer/

Menü

AUTO SCOUT 24

Zurück

Sortieren

Dino

8 Suchergebnisse für Fiat Dino

Fiat Dino ×

Deutschland ×

Speichern

Benachrichtigung

Filtern

1/15

Fiat Dino

Spider 2,0 / VIN 0078!!

Merken

Teilen

€ 137.900,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

Aston Martin vor 1970

31 Suchergebnisse für Aston Martin

Aston Martin ×

Erstzulassung bis zu 1970 ×

Speichern

Benachrichtigung

Filtern

1/15

Aston Martin DB

DB2/4 MK III

Merken

Teilen

€ 225.000,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

Rotes Coupe 1960-1980

306 Suchergebnisse

Erstzulassung von 1960 ×

Erstzulassung bis 1980 ×

Speichern

Benachrichtigung

Filtern

1/10

Plymouth Road Runner

Rennwagen Hist. NASCAR

Merken

Teilen

€ 14.500,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

396 cui

1 Suchergebnis

Variante: 396 cui ×

Deutschland ×

Speichern

Benachrichtigung

Filtern

Chevrolet El Camino

SS 396 cui V8 H-Zulassung

Merken

Teilen

€ 31.500,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

Ford V8 vor 1973

164 Suchergebnisse für Ford v8

Ford ×

Variante: v8 ×

Erstzulassung bis zu 1973

Speichern

Benachrichtigung

Filtern

Ford Maverick

Grabber, 302ci V8, Top-Restau…

Merken

Teilen

€ 26.900,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

vor 1980 über 200 ps 30.000-50.000€

364 Suchergebnisse

Verkaufspreis von 30.000 € ×

Verkaufspreis bis 50.

Speichern

Benachrichtigung

Filtern

Pontiac

Catalina

Merken

Teilen

€ 30.000,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

BMW 2002

39 Suchergebnisse für BMW 2002

BMW 2002 ×

Deutschland ×

Speichern

Benachrichtigung

Filtern

BMW 2002

2000,erst 66Tkm,1.Hd,Erstlac…

Merken

Teilen

€ 22.500,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

BMW vor 2002

6.641 Suchergebnisse für BMW

BMW ×

Erstzulassung bis zu 2002 ×

Deutschland

Speichern

Benachrichtigung

Filtern

BMW 635

6er M6 M635CSI

Merken

Teilen

€ 18.500,-

KEINE ANGABE

Menü

AUTO SCOUT 24

Zurück

Sortieren

BMW unter 2002 eur

2.267 Suchergebnisse für BMW

BMW ×

Verkaufspreis bis zu 2.000 € ×

Deutschla

Speichern

Benachrichtigung

Filtern

BMW 318

3er

Merken

Teilen

€ 1.150,-

GUTES ANGEBOT

VARIETY IN NATURAL LANGUAGE

peogot
peogout
peuceot
peugeaut
peugeot
peugeu
peugoet
peugot
peujeot
peuscho

VARIETY IN NATURAL LANGUAGE

peogot
peogout
peuceot
peugeaut
peugeot
peugeu
peugoet
peugot
peujeot
peuscho

renalt
renaud
renaul
renauld
renault
renaults
renaut
renolte
renoult
rinault

VARIETY IN NATURAL LANGUAGE

peogot
peogout
peuceot
peugeaut
peugeot
peugeu
peugoet
peugot
peujeot
peuscho

renalt
renaud
renaul
renauld
renault
renaults
renaut
renolte
renoult
rinault

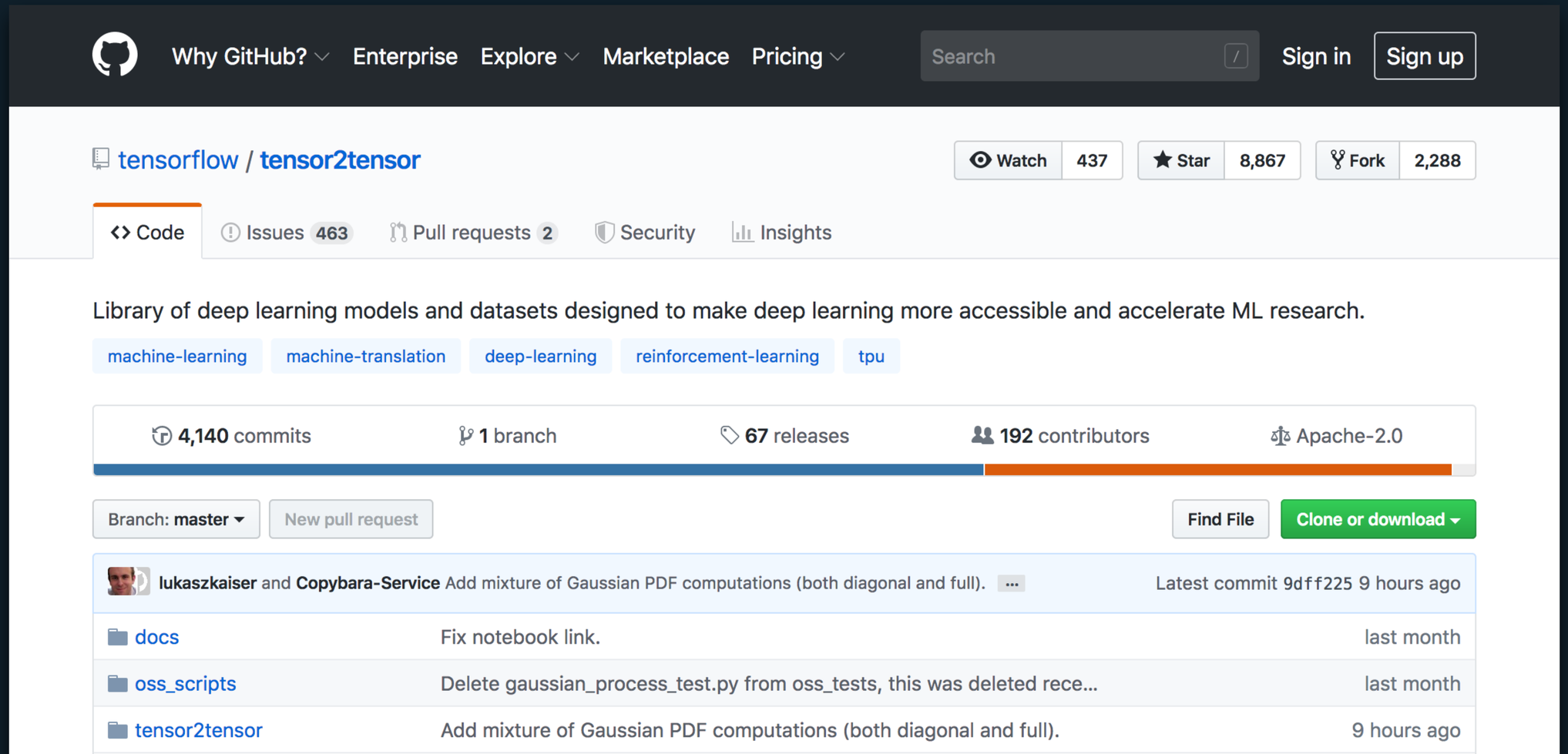
volcwagen
volfsvagen
volksagen
volkswagem
volkswagen
…
wolcvagen
wolksvagen
wolsfagen
wolsvagan

DATA

“In God we trust, all others bring data.”

–W. EDWARDS DEMING

TRAINING


Why GitHub?
Enterprise
Explore
Marketplace
Pricing
Search
/
Sign in
Sign up
tensorflow / tensor2tensor
Watch 437
Star 8,867
Fork 2,288
Code
Issues 463
Pull requests 2
Security
Insights
Library of deep learning models and datasets designed to make deep learning more accessible and accelerate ML research.
machine-learning
machine-translation
deep-learning
reinforcement-learning
tpu
4,140 commits
1 branch
67 releases
192 contributors
Apache-2.0
Branch: master
New pull request
Find File
Clone or download
lukaszkaiser and Copybara-Service Add mixture of Gaussian PDF computations (both diagonal and full).
Latest commit 9dff225 9 hours ago
docs
Fix notebook link.
last month
oss_scripts
Delete gaussian_process_test.py from oss_tests, this was deleted rece...
last month
tensor2tensor
Add mixture of Gaussian PDF computations (both diagonal and full).
9 hours ago

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Published as a conference paper at ICLR 2019

UNIVERSAL TRANSFORMERS

Mostafa Dehghani*† University of Amsterdam dehghani@uva.nl
Stephan Gouws* DeepMind sgouws@google.com
Oriol Vinyals DeepMind vinyals@google.com
Jakob Uszkoreit Google Brain usz@google.com
Łukasz Kaiser Google Brain lukaszkaiser@google.com

ABSTRACT

Recurrent neural networks (RNNs) sequentially process data by updating their state with each new data point, and have long been the de facto choice for sequence modeling tasks. However, their inherently sequential computation makes them slow to train. Feed-forward and convolutional architectures have recently been shown to achieve superior results on some sequence modeling tasks such as machine translation, with the added advantage that they concurrently process all inputs in the sequence, leading to easy parallelization and faster training times. Despite these successes, however, popular feed-forward sequence models like the Transformer fail to generalize in many simple tasks that recurrent models handle with ease, e.g. copying strings or even simple logical inference when the string or formula lengths exceed those observed at training time. We propose the Universal Transformer (UT), a parallel-in-time self-attentive recurrent sequence model which can be cast as a generalization of the Transformer model and which addresses these issues. UTs combine the parallelizability and global receptive field of feed-forward sequence models like the Transformer with the recurrent inductive bias of RNNs. We also add a dynamic per-position halting mechanism and find that it improves accuracy on several tasks. In contrast to the standard Transformer, under certain assumptions UTs can be shown to be Turing-complete. Our experiments show that UTs outperform standard Transformers on a wide range of algorithmic and language understanding tasks, including the challenging LAMBADA language modeling task where UTs achieve a new state of the art, and machine translation where UTs achieve a 0.9 BLEU improvement over Transformers on the WMT14 En-De dataset.

1 INTRODUCTION

Fast Decoding in Sequence Models Using Discrete Latent Variables

Łukasz Kaiser [1] **Aurko Roy** [1] **Ashish Vaswani** [1] **Niki Parmar** [1] **Samy Bengio** [1] **Jakob Uszkoreit** [1] **Noam Shazeer** [1]

Abstract

Autoregressive sequence models based on deep neural networks, such as RNNs, Wavenet and the Transformer attain state-of-the-art results on many tasks. However, they are difficult to parallelize and are thus slow at processing long sequences. RNNs lack parallelism both during training and decoding, while architectures like WaveNet and Transformer are much more parallelizable during training, yet still operate sequentially during decoding. We present a method to extend sequence models using discrete latent variables that makes decoding much more parallelizable. We first autoencode the target sequence into a shorter sequence of discrete latent variables, which at inference time is generated autoregressively, and finally decode the output sequence from this shorter latent sequence in parallel. To this end, we introduce a novel method for constructing a sequence of discrete latent variables and compare it with previously introduced methods. Finally, we evaluate our model end-to-end on the task of neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014), parsing (Vinyals et al., 2015), and many others. RNNs are inherently sequential, however, and thus tend to be slow to execute on modern hardware optimized for parallel execution. Recently, a number of more parallelizable sequence models were proposed and architectures such as WaveNet (van den Oord et al., 2016), ByteNet (Kalchbrenner et al., 2016) and the Transformer (Vaswani et al., 2017) can indeed be trained faster due to improved parallelism.

When actually generating sequential output, however, their autoregressive nature still fundamentally prevents these models from taking full advantage of parallel computation. When generating a sequence $y_1 \dots y_n$ in a canonical order, say from left to right, predicting the symbol y_t first requires generating all symbols $y_1 \dots y_{t-1}$ as the model predicts

$$P(y_t|y_{t-1}\ y_{t-2} \cdots\ y_1).$$

During training, the ground truth is known so the conditioning on previous symbols can be parallelized. But during decoding, this is a fundamental limitation as at least n sequential steps need to be made to generate $y_1 \dots y_n$.

1. I

Neu
of ta
rent
ory (
sequ
have

[1]
denc
Roy

Proc
Lear
by th

The Evolved Transformer

David R. So [1] **Chen Liang** [1] **Quoc V. Le** [1]

Abstract

Recent works have highlighted the strength of the *Transformer* architecture on sequence tasks while, at the same time, *neural architecture search* (NAS) has begun to outperform human-designed models. Our goal is to apply NAS to search for a better alternative to the Transformer. We first construct a large search space inspired by the recent advances in feed-forward sequence models and then run evolutionary architecture search with *warm starting* by seeding our initial population with the Transformer. To directly search on the computationally expensive WMT 2014 English-German translation task, we develop the *Progressive Dynamic Hurdles* method, which allows us to dynamically allocate more resources to more promising candidate models. The architecture found in our experiments – the *Evolved Transformer* – demonstrates consistent improvement over the Transformer on four well-established language tasks: WMT 2014 English-German, WMT 2014 English-French, WMT 2014 English-Czech and LM1B. At a big model size, the Evolved Transformer establishes a new state-of-the-art BLEU score of 29.8 on WMT'14 English-German; at smaller sizes, it achieves the same quality as the original "big" Transformer with 37.6% less parameters and outperforms the Transformer by 0.7 BLEU at a mobile-friendly model

models, although some effort has also been invested in searching for sequence models (Zoph & Le, 2017; Pham et al., 2018). In these cases, it has always been to find improved recurrent neural networks (RNNs), which were long established as the de facto neural model for sequence problems (Sutskever et al., 2014; Bahdanau et al., 2015).

However, recent works have shown that there are better alternatives to RNNs for solving sequence problems. Due to the success of convolution-based networks, such as Convolution Seq2Seq (Gehring et al., 2017), and full attention networks, such as the Transformer (Vaswani et al., 2017), feed-forward networks are now a viable option for solving sequence-to-sequence (seq2seq) tasks. The main strength of feed-forward networks is that they are faster, and easier to train than RNNs.

The goal of this work is to examine the use of neural architecture search methods to design better feed-forward architectures for seq2seq tasks. Specifically, we apply *tournament selection* architecture search and *warm start* it with the Transformer, considered to be the state-of-art and widely-used, to evolve a better and more efficient architecture. To achieve this, we construct a search space that reflects the recent advances in feed-forward seq2seq models and develop a method called *Progressive Dynamic Hurdles* (PDH) that allows us to perform our search directly on the computationally demanding WMT 2014 English-German (En-De) translation task. Our search produces a new architecture – called the *Evolved Transformer* (ET) –

```
pip install tensor2tensor

t2t-trainer \
  --generate_data \
  --model=$MODEL \
  --problem=$PROBLEM \
  --data_dir=$DATA_DIR \
  --hparams_set=$HPARAMS \
  --output_dir=$TRAIN_DIR
```

TensorBoard SCALARS IMAGES GRAPHS PROJECTOR TEXT INACTIVE

Show data download links

Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing

0

Horizontal Axis

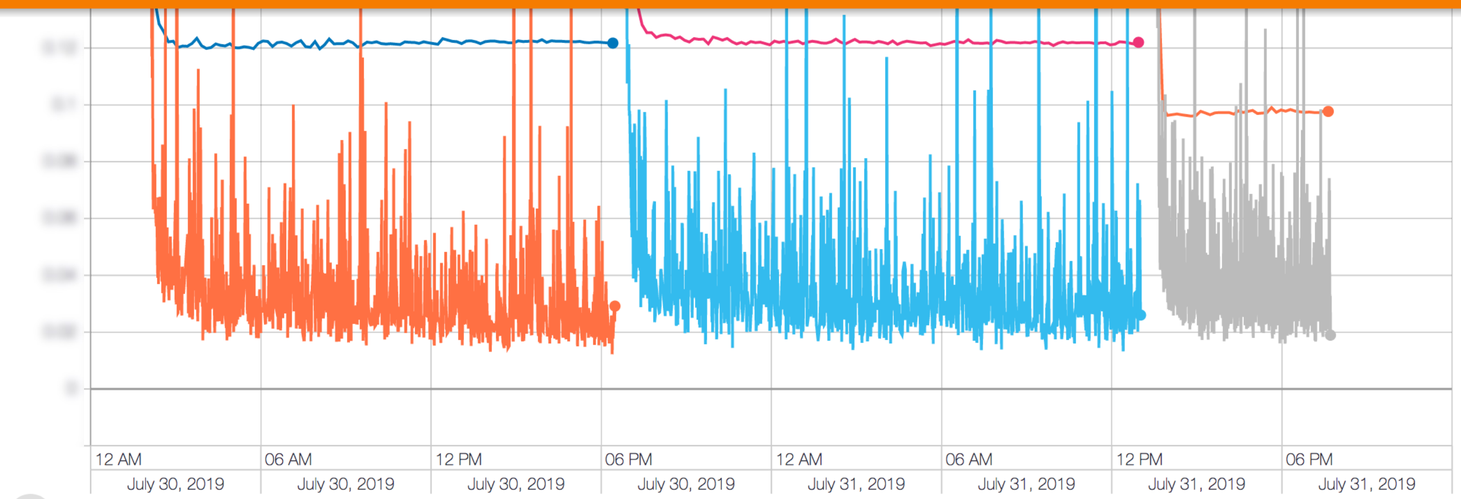
STEP RELATIVE WALL

Runs

Write a regex to filter runs

TOGGLE ALL RUNS

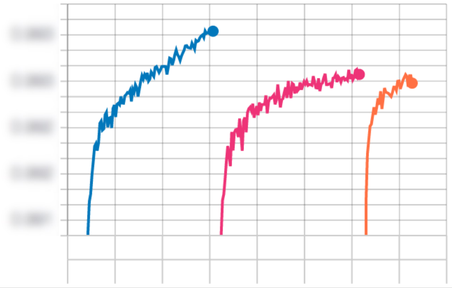
/ec2/train



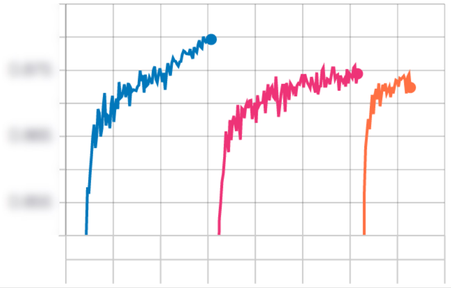
losses 2

metrics-translate_query 7

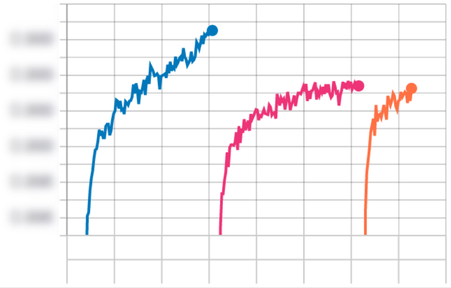
targets/accuracy
tag: metrics-translate_query/targets/accuracy



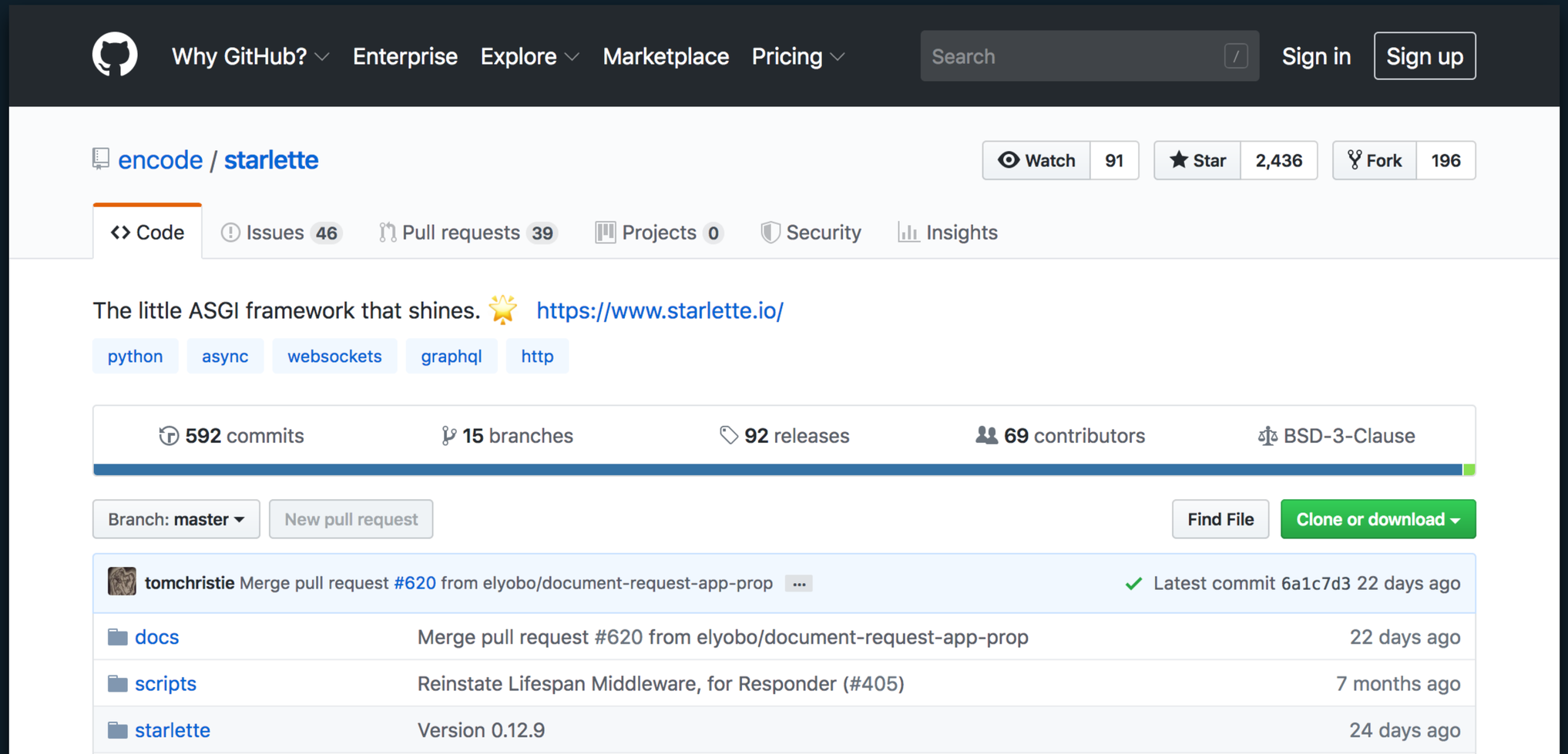
targets/accuracy_per_sequence
tag: metrics-translate_query/targets/accuracy_per_sequence



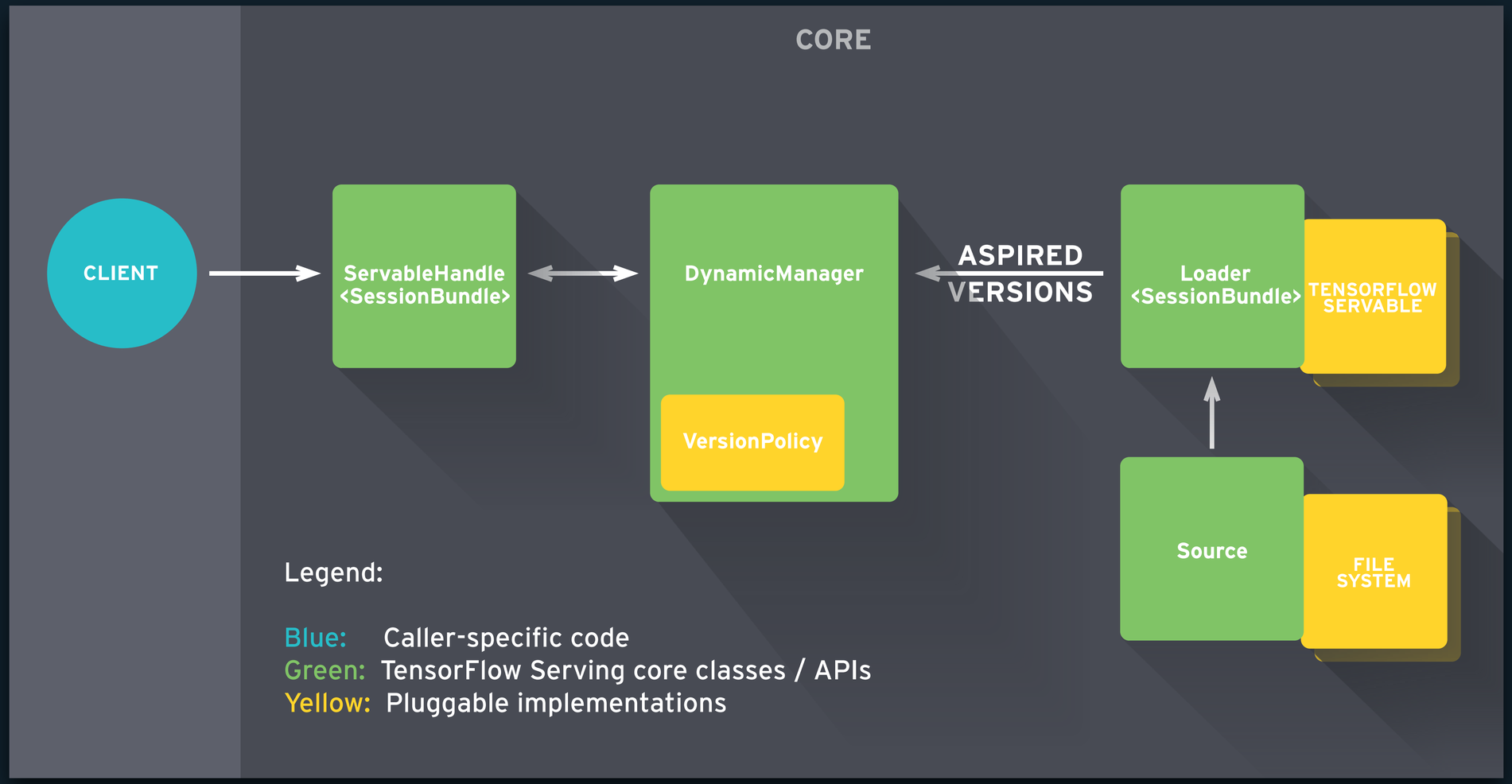
targets/accuracy_top5
tag: metrics-translate_query/targets/accuracy_top5



DEPLOYMENT API



DEPLOYMENT MODEL



Source: https://www.tensorflow.org/tfx/serving/architecture

THANK YOU FOR YOUR ATTENTION!

https://www.linkedin.com/in/markusludwig/