GoogleResearch

Research Areas: Natural Language Processing, Responsible AI, Speech Processing
Years: 2023, 2022, 2021, 2020, 2019, 2018, 2017
Number of Research Papers Evaluated: 1,138
Number of Research Papers After Evaluation: 143
Author: Dhruv Awasthi

(QA)^2: Question Answering with Questionable Assumptions (ACL 2023; Link)

Abstract:
Naturally-occurring information-seeking questions often contain questionable assumptions---assumptions that are false or unverifiable. Questions containing questionable assumptions are challenging because they require a distinct answer strategy that deviates from typical answers to information-seeking questions. For instance, the question "When did Marie Curie discover Uranium?" cannot be answered as a typical "when" question without addressing the false assumption "Marie Curie discovered Uranium". In this work, we propose (QA)^2 (Question Answering with Questionable Assumptions), an open-domain evaluation dataset consisting of naturally-occurring search engine queries that may or may not contain questionable assumptions. To be successful on (QA)^2, systems must be able to detect questionable assumptions and also be able to produce adequate responses for both typical information-seeking questions and ones with questionable assumptions. Through human rater acceptability on abstractive QA with (QA)^2 questions, we find that current models do struggle with handling questionable assumptions, leaving substantial headroom for progress.

Coreference Resolution through a seq2seq Transition-Based System (TACL 2023; Link)

Abstract:
Most recent coreference resolution systems use search algorithms over possible spans to identify mentions and resolve coreference. We instead present a coreference resolution system that uses a text-to-text (seq2seq) paradigm to predict mentions and links jointly, which simplifies the coreference resolution by eliminating both the search for mentions and coreferences. We implemented the coreference system as a transition system and use multilingual T5 as language model. We obtained state-of-the-art accuracy with 83.3 F1-score on the CoNLL-2012 data set. We use the SemEval-2010 data sets to evaluate on languages other than English and get substantially higher Zero-shot F1-scores for 3 out of 4 languages

than previous approaches and significantly exceed previous supervised state-of-the-art results for all five tested languages.

Covering Uncommon Ground: Followup Question Generation for Answer Assessment (ACL 2023; Link)

Abstract:
In educational dialogue settings students often provide answers that are incomplete. In other words, there is a gap between the answer the student provides and the perfect answer expected by the teacher. Successful dialogue hinges on the teacher asking about this gap in an effective manner, thus creating a rich and interactive educational experience. Here we focus on the problem of generating such gap-focused questions (GFQ) automatically. We define the task, highlight key desired aspects of a good GFQ, and propose a model that satisfies these. Finally, we provide an evaluation of our generated questions and compare them to manually generated ones, demonstrating competitive performance.

Discriminative Diffusion Models as Few-shot Vision and Language Learners (ArXiv 2023; Link)

Abstract:
Diffusion models, such as Stable Diffusion, have shown incredible performance on text-to-image generation. Since text-to-image generation often requires models to generate visual concepts with fine-grained details and attributes specified in text prompts, can we leverage the powerful representations learned by pre-trained diffusion models for discriminative tasks such as image-text matching? To answer this question, we propose a novel approach, Discriminative Stable Diffusion (DSD), which turns pre-trained text-to-image diffusion models into few-shot discriminative learners. Our approach uses the cross-attention score of a Stable Diffusion model to capture the mutual influence between visual and textual information and fine-tune the model via attention-based prompt learning to perform image-text matching. By comparing DSD with state-of-the-art methods on several benchmark datasets, we demonstrate the potential of using pre-trained diffusion models for discriminative tasks with superior results on few-shot image-text matching.

Disentangling speech from surroundings with neural embeddings (ICASSP 2023; Link)

Abstract:
We present a method to separate speech signals from noisy environments in the embedding space of a neural audio codec. We introduce a new training procedure that allows our model to produce structured encodings of audio waveforms given by embedding vectors, where one part of the embedding vector represents the speech signal, and the rest represent the environment. We achieve this by partitioning the embeddings of different input waveforms

and training the model to faithfully reconstruct audio from mixed partitions, thereby ensuring each partition encodes a separate audio attribute. As use cases, we demonstrate the separation of speech from background noise or from reverberation characteristics. Our method also allows for targeted adjustments of the audio output characteristics.

Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes (ACL 2023; Link)

Abstract:
Deploying large language models (LLMs) is challenging because they are memory inefficient and compute-intensive for practical applications. In reaction, researchers train smaller task-specific models by either finetuning with human labels or distilling using LLM-generated labels. However, finetuning and distillation require large amounts of training data to achieve comparable performance to LLMs. We introduce Distilling step-by-step, a new mechanism that (a) trains smaller models that outperform LLMs, and (b) achieves so by leveraging less training data needed by finetuning or distillation. Our method extracts LLM rationales as additional supervision for small models within a multi-task training framework. We present three findings across 4 NLP benchmarks: First, compared to both finetuning and distillation, our mechanism achieves better performance with much fewer labeled/unlabeled training examples. Second, compared to LLMs, we achieve better performance using substantially smaller model sizes. Third, we reduce both the model size and the amount of data required to outperform LLMs; our 770M T5 model outperforms the 540B PaLM model using only 80% of available data on a benchmark task.

Diversifying Joint Vision-Language Tokenization Learning (CVPR 2023; Link)

Abstract:
Building joint representations across images and text is an essential step for tasks such as Visual Question Answering and Video Question Answering. In this work, we find that the representations must not only jointly capture features from both modalities but should also be diverse for better generalization performance. To this end, we propose joint vision-language representation learning by diversifying the tokenization learning process, enabling tokens which are sufficiently disentangled from each other to be learned from both modalities. We observe that our approach outperforms the baseline models in a majority of settings and is competitive with state-of-the-art methods.

Envisioning Equitable Speech Technologies for Black Older Adults (FaCCT 2023; Link)

Abstract:
There is increasing concern that how researchers currently define and measure fairness is inadequate. Recent calls push to move beyond traditional concepts of fairness and consider related constructs through qualitative and community-based approaches, particularly for underrepresented communities most at-risk for AI harm. One in context, previous research has identified that voice technologies are unfair due to racial and age disparities. This paper uses voice technologies as a case study to unpack how Black older adults value and envision fair and equitable AI systems. We conducted design workshops and interviews with 16 Black older adults, exploring how participants envisioned voice technologies that better understand cultural context and mitigate cultural dissonance. Our findings identify tensions between what it means to have fair, inclusive, and representative voice technologies. This research raises questions about how and whether researchers can model cultural representation with large language models.

Extracting Representative Subset from Massive Raw Texts for Training Pre-trained Neural Language Models (Information Processing & Management Conference 2023; Link)

Abstract:
This paper explores the research question of whether training neural language models using a small subset of representative data selected from a large training dataset can achieve the same level of performance obtained using all the original training data. We explore the likelihood-based scoring for the purpose of obtaining representative subsets, which we call RepSet. Our experiments confirm that the representative subset obtained by a likelihood difference-based score can achieve the 90% performance level even when the dataset is reduced to about 1,000th of the original data. We also show that the performance of the random selection method deteriorates significantly when the amount of data is reduced.

Generative Information Retrieval (ACM SIGIR 2023; Link)

Abstract:
Historically, information retrieval systems have all followed the same paradigm: information seekers frame their needs in the form of a short query, the system selects a small set of relevant results from a corpus of available documents, rank-orders the results by decreasing relevance, possibly excerpts a responsive passage for each result, and returns a list of references and excerpts to the user. Retrieval systems typically did not attempt fusing information from multiple documents into an answer and displaying that answer directly. This was largely due to available technology: at the core of each retrieval system is an index that

maps lexical tokens or semantic embeddings to document identifiers. Indices are designed for retrieving responsive documents; they do not support integrating these documents into a holistic answer.

More recently, the coming-of-age of deep neural networks has dramatically improved the capabilities of large language models (LLMs). Trained on a large corpus of documents, these models not only memorize the vocabulary, morphology and syntax of human languages, but have shown to be able to memorize facts and relations [2]. Generative language models, when provided with a prompt, will extend the prompt with likely completions – an ability that can be used to extract answers to questions from the model. Two years ago, Metzler et al. argued that this ability of LLMs will allow us to rethink the search paradigm: to answer information needs directly rather that directing users to responsive primary sources [1]. Their vision was not without controversy; the following year Shaw and Bender argued that such a system is neither feasible nor desirable [3]. Nonetheless, the past year has seen the emergence of such systems, with offerings from established search engines and multiple new entrants to the industry.

The keynote will summarize the short history of these generative information retrieval systems, and focus on the many open challenges in this emerging field: ensuring that answers are grounded, attributing answer passages to a primary source, providing nuanced answers to non-factoid-seeking questions, avoiding bias, and going beyond simple regurgitation of memorized facts. It will also touch on the changing nature of the content ecosystem. LLMs are starting to be used to generate web content. Should search engines treat such derived content equal to human-authored content? Is it possible to distinguish generated from original content? How should we view hybrid authorship where humans contribute ideas and LLMs shape these ideas into prose? And how will this parallel technical evolution of search engines and content ecosystems affect their respective business models?

Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning (2023; Link)

Abstract:
Language models still struggle on moral reasoning, despite their impressive performance in many other tasks. In particular, the Moral Scenarios task in MMLU (Multi-task Language Understanding) is among the worst performing tasks for many language models, including GPT-3. In this work, we propose a new prompting framework, Thought Experiments, to teach language models to do better moral reasoning using counterfactuals. Experiment results show that our framework elicits counterfactual questions and answers from the model, which in turn helps improve the accuracy on Moral Scenarios task by 9-16% compared to other zero-shot baselines. Interestingly, unlike math reasoning tasks, zero-shot Chain-of-Thought (CoT) reasoning doesn't work out of the box, and even reduces accuracy by around 4% compared to direct zero-shot. We further observed that with minimal human supervision in the form of 5 few-shot examples, the accuracy of the task can be improved to as much as 80%.

MaXM: Towards Multilingual Visual Question Answering (EMNLP 2023; Link)

Abstract:
Visual Question Answering (VQA) has been primarily studied through the lens of the English language. Yet, tackling VQA in other languages in the same manner would require a considerable amount of resources. In this paper, we propose scalable solutions to multilingual visual question answering (mVQA), on both data and modeling fronts. We first propose a translation-based framework to mVQA data generation that requires much less human annotation efforts than the conventional approach of directly collection questions and answers. Then, we apply our framework to the multilingual captions in the Crossmodal-3600 dataset and develop an efficient annotation protocol to create MaXM, a test-only VQA benchmark in 7 diverse languages. Finally, we develop a simple, lightweight, and effective approach as well as benchmark state-of-the-art English and multilingual VQA models. We hope that our benchmark encourages further research on mVQA.

MISGENDERED: Limits of Large Language Models in Understanding Pronouns (ACL 2023; Link)

Abstract:
Gender bias in language technologies has been widely studied, but research has mostly been restricted to a binary paradigm of gender. It is important to also consider non-binary gender identities, as excluding them can cause further harm to an already marginalized group. One way in which English-speaking individuals linguistically encode their gender identity is through third-person personal pronoun declarations. This is often done using two or more pronoun forms, e.g., \textit{xe/xem}, or \textit{xe/xem/xyr}. In this paper, we comprehensively evaluate state-of-the-art language models for their ability to correctly use declared third-person personal pronouns. As far as we are aware, we are the first to do so. We evaluate language models in both zero-shot and few-shot settings. Models are still far from zero-shot gendering non-binary individuals accurately, and most also struggle with correctly using gender-neutral pronouns (singular \textit{they, them, their} etc.). This poor performance may be due to the lack of representation of non-binary pronouns in pre-training corpora, and some memorized associations between pronouns and names. We find an overall improvement in performance for non-binary pronouns when using in-context learning, demonstrating that language models with few-shot capabilities can adapt to using declared pronouns correctly.

MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering (CVPR 2023; Link)

Abstract:
To build Video Question Answering (VideoQA) systems capable of assisting humans in daily activities, seeking answers from long-form videos with diverse and complex events is a must. Existing multi-modal VQA models achieve promising performance on images or short video clips, especially with the recent success of large-scale multi-modal pre-training. However, when extending these methods to long-form videos, new challenges arise. On the one hand, using a dense video sampling strategy is computationally prohibitive. On the other hand, methods relying on sparse sampling struggle in scenarios where multi-event and multi-granularity visual reasoning are required. In this work, we introduce a new model named Multi-modal Iterative Spatial-temporal Transformer (MIST) to better adapt pre-trained models for long-form VideoQA. Specifically, MIST decomposes traditional dense spatial-temporal self-attention into cascaded segment and region selection modules that adaptively select frames and image regions that are closely relevant to the question itself. Visual concepts at different granularities are then processed efficiently through an attention module. In addition, MIST iteratively conducts selection and attention over multiple layers to support reasoning over multiple events. The experimental results on four VideoQA datasets, including AGQA, NExT-QA, STAR, and Env-QA, show that MIST achieves state-of-the-art performance and is superior at computation efficiency and interpretability.

Modular Domain Adaptation for Conformer-Based Streaming ASR (Interspeech 2023; Link)

Abstract:
Speech data from different domains has distinct acoustic and linguistic characteristics. It is common to train a single multidomain model such as a Conformer transducer for speech recognition on a mixture of data from all domains. However, changing data in one domain or adding a new domain would require the multidomain model to be retrained. To this end, we propose a framework called modular domain adaptation (MDA) that enables a single model to process multidomain data while keeping all parameters domain-specific, i.e., each parameter is only trained by data from one domain. On a streaming Conformer transducer trained only on video caption data, experimental results show that an MDA-based model can reach similar performance as the multidomain model on other domains such as voice search and dictation by adding per-domain adapters and per-domain feed-forward networks in the Conformer encoder.

MoQA: Benchmarking Multi-Type Open-Domain Question Answering (ACL 2023; Link)

Abstract:
Existing open-domain question answering research mainly focuses on questions that can be answered in a few words. However, information-seeking questions often require different formats of answers depending on the nature of questions, e.g., ``Why is there a maple leaf on the Canadian flag?'' In this paper, we present a new task, MOQA, which requires building QA models that can provide short, medium, long, and yes/no answers to open-domain questions simultaneously. We expand the Natural Questions dataset into the open-domain setting by keeping all types of questions and show that existing systems cannot generalize to these new types. We adapt state-of-the-art open-domain QA models---based on retriever-reader and phrase retrieval models---to tackle this task. Results and analyses of our multi-type QA models reveal the unique challenges of the task, calling for versatile QA models in the future.

OpineSum: Entailment-based self-training for abstractive opinion summarization (ACL 2023; Link)

Abstract:
A typical product or place often has hundreds of reviews, and summarization of these texts is an important and challenging problem. Recent progress on abstractive summarization in domains such as news has been driven by supervised systems trained on hundreds of thousands of news articles paired with human-written summaries. However for opinion texts, such large scale datasets are rarely available. Unsupervised methods, self-training, and few-shot learning approaches bridge that gap. In this work, we present a novel self-training approach, OpineSum for abstractive opinion summarization. The summaries in this approach are built using a novel application of textual entailment and capture the consensus of opinions across the various reviews for an item. This method can be used to obtain silver-standard summaries on a large scale and train both unsupervised and few-shot abstractive summarization systems. OpineSum achieves state-of-the-art performance in both settings.

PaLI: A Jointly-Scaled Multilingual Language-Image Model (ICLR 2023; Link)

Abstract:
Effective scaling and a flexible task interface enable large-capacity language models to excel at many tasks. PaLI (Pathways Language and Image model) extends these ideas to the joint modeling of language and vision. PaLI is a model that generates text based on visual and textual inputs. Using this API, PaLI is able to perform many vision, language, and multimodal tasks, across many languages. We train PaLI with two main principles: reuse of pretrained unimodal components, and joint scaling of modalities. Using large-capacity pretrained language models and vision models allows us to capitalize on their existing capabilities, while leveraging the substantial cost of training them. We scale PaLI models across three

axes:the language component, the vision component, and the training data that fuses them. For the vision component, we train the largest and best-performing VisionTransformer (ViT) to date. For the data, we build an image-text training set over10B images and covering over 100 languages. PaLI inherits and enhances language-understanding capabilities, and achieves state-of-the-art in multiple vision and language tasks (image classification, image captioning, visual question-answering, scene-text understanding, etc.), based on a simple, modular, and reuse-friendly platform for modeling and scaling.

Parameter-Efficient Finetuning for Robust Continual Multilingual Learning (ACL 2023; Link)

Abstract:
We introduce and study the problem of Continual Multilingual Learning (CML), where a previously trained multilingual model is periodically updated using new data arriving in stages. If the new data is present only in a subset of languages, we find that the resulting model shows improved performance only on the languages included in the latest update (and few closely related languages) while its performance on all the remaining languages degrade significantly. We address this challenge by proposing LAFT-URIEL, a parameter-efficient finetuning strategy which aims to increase the number of languages on which the model improves after an update, while reducing the magnitude of loss in performance for the remaining languages. LAFT-URIEL uses linguistic knowledge to balance overfitting and knowledge sharing across languages, thus resulting in 25% increase in the number of languages whose performances improve during an update and 78% relative decrease in average magnitude of losses on the remaining languages.

PreSTU: Pre-Training for Scene-Text Understanding (ICCV 2023; Link)

Abstract:
The ability to recognize and reason about text embedded in visual inputs is often lacking in vision-and-language (V&L) models, perhaps because V&L pre-training methods have often failed to include such an ability in their training objective. In this paper, we propose PreSTU, a novel pre-training recipe dedicated to scene-text understanding (STU). PreSTU introduces OCR-aware pre-training objectives that encourage the model to recognize text from an image and connect it to the rest of the image content. We implement PreSTU using a simple transformer-based encoder-decoder architecture, combined with large-scale image-text datasets with scene text obtained from an off-the-shelf OCR system. We empirically demonstrate the effectiveness of this pre-training approach on eight visual question answering and four image captioning benchmarks.

QueryForm: A Simple Zero-shot Form Entity Query Framework (ACL 2023; Link)

Abstract:
Zero-shot transfer learning for document understanding is a crucial yet under-investigated scenario to help reduce the high cost involved in annotating document entities. We present a novel query-based framework, QueryForm, that extracts entity values from form-like documents in a zero-shot fashion. QueryForm contains a dual prompting mechanism that composes both the document schema and a specific entity type into a query, which is used to prompt a Transformer model to perform a single entity extraction task. Furthermore, we propose to leverage large-scale query-entity pairs generated from form-like webpages with weak HTML annotations to pre-train QueryForm. By unifying pre-training and fine-tuning into the same query-based framework, QueryForm enables models to learn from structured documents containing various entities and layouts, leading to better generalization to target document types without the need for target-specific training data. QueryForm sets new state-of-the-art average F1 score on both the XFUND (+4.6%~10.1%) and the Payment (+3.2%~9.5%) zero-shot benchmark, with a smaller model size and no additional image input.

RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses (SIGIR 2023; Link)

Abstract:
Pretrained language models such as BERT have been shown to be exceptionally effective for text ranking. However, there are limited studies on how to leverage more powerful sequence-to-sequence models such as T5. Existing attempts usually formulate text ranking as a classification problem and rely on postprocessing to obtain a ranked list. In this paper, we propose RankT5 and study two T5-based ranking model structures, an encoder-decoder and an encoder-only one, so that they not only can directly output ranking scores for each query-document pair, but also can be fine-tuned with "pairwise" or "listwise" ranking losses to optimize ranking performance. Our experiments show that the proposed models with ranking losses can achieve substantial ranking performance gains on different public text ranking data sets. Moreover, ranking models fine-tuned with listwise ranking losses have better zero-shot ranking performance on out-of-domain data than models fine-tuned with classification losses.

Recitation-Augmented Language Models (ICLR 2023; Link)

Abstract:
We propose a new paradigm to help Large Language Models (LLMs) generate more accurate factual knowledge without retrieving from an external corpus, called

RECITation-augmented gEneration (RECITE). Different from retrieval-augmented language models that retrieve relevant documents before generating the outputs, given an input, RECITE first recites one or several relevant passages from LLMs' own memory via sampling, and then produces the final answers. We show that RECITE is a powerful paradigm for knowledge-intensive NLP tasks. Specifically, we show that by utilizing recitation as the intermediate step, a recite-and-answer scheme can achieve new state-of-the-art performance in various closed-book question answering (CBQA) tasks. In experiments, we verify the effectiveness of RECITE on three pre-trained models (PaLM, UL2, and OPT) and three CBQA tasks (Natural Questions, TriviaQA, and HotpotQA).

Tool Documentation Enables Zero-Shot Tool-Usage with Large Language Models (2023; Link)

Abstract:
Today, large language models (LLMs) are taught to use new tools by providing a few demonstrations of the tool's usage. Unfortunately, demonstrations are hard to acquire, and can result in undesirable biased usage if the wrong demonstration is chosen. Even in the rare scenario that demonstrations are readily available, there is no principled selection protocol to determine how many and which ones to provide. As tasks grow more complex, the selection search grows combinatorially and invariably becomes intractable. Our work provides an alternative to \textbf{demonstrations}: tool textbf{documentation}. We advocate the use of tool documentation—descriptions for the individual tool usage—over demonstrations. We substantiate our claim through three main empirical findings on 6 tasks across both vision and language modalities. First, on existing benchmarks, zero-shot prompts with only tool documentation are sufficient for eliciting proper tool usage, achieving performance on par with few-shot prompts. Second, on a newly collected realistic tool-use dataset with hundreds of available tool APIs, we show that tool documentation is significantly more valuable than demonstrations, with zero-shot documentation significantly outperforming few-shot without documentation. Third, we highlight the benefits of tool documentations by tackling image generation and video tracking using just-released unseen state-of-the-art models as tools. Finally, we highlight the possibility of using tool documentation to automatically enable new applications: by using nothing more than the documentation of GroundingDino, Stable Diffusion, XMem, and SAM, LLMs can \emph{re-invent} the functionalities of the just-released Grounded-SAM~\cite{groundedsam} and Track Anything~\cite{yang2023track} models.

Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters (ACL 2023; Link)

Abstract:
Chain-of-Thought (CoT) prompting can dramatically improve the multi-step reasoning abilities of large language models (LLMs). CoT explicitly encourages the LLM to generate

intermediate rationales for solving a problem, by providing a series of reasoning steps in the demonstrations. Despite its success, there is still little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. In this paper, we show that CoT reasoning is possible even with invalid demonstrations - prompting with invalid reasoning steps can achieve over 80-90% of the performance obtained using CoT under various metrics, while still generating coherent lines of reasoning during inference. Further experiments show that other aspects of the rationales, such as being relevant to the query and correctly ordering the reasoning steps, are much more important for effective CoT reasoning. Overall, these findings both deepen our understanding of CoT prompting, and open up new questions regarding LLMs' capability to learn to reason in context.

Transferring Visual Attributes from Natural Language to Verified Image Generation (2023; Link)

Abstract:
Text to image generation methods (T2I) are widely popular in generating art and other creative artifacts. While hallucination can be a positive factor in scenarios where creativity is appreciated, such artifacts are poorly suited for tasks where the generated image needs to be grounded in a strict manner, e.g. as an illustration of a task, an action or in the context of a story. In this paper, we propose to strengthen the factual consistency properties of T2I methods in the presence of natural prompts. First, we cast the problem as an MT problem that translates natural prompts into visual prompts. Then we filter the image with a VQA approach where we answer a set of questions in the visual domain (the image) and in the natural language domain (the natural prompt). Finally, to measure the alignment of answers, we depart from the recent literature that do string matching, and compare answers in an embedding space that assesses the semantic and entailment associations between a natural prompt and its generated image.

What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis (ACM Web Conference 2023; Link)

Abstract:
Market sentiment analysis on social media content requires knowledge of both financial markets and social media slang, which makes it a challenging task for human raters. The resulting lack of high-quality labeled data stands in the way of conventional supervised learning methods. Instead, we approach this problem using semi-supervised learning with a large language model (LLM). Our pipeline generates weak financial sentiment labels for Reddit posts with an LLM and then uses that data to train a small model that can be served in production. We find that prompting the LLM to produce chain-of-thought summaries and forcing it through several reasoning paths helps generate more stable and accurate labels, while using a regression loss further improves distillation quality. With only a handful of

prompts, the final model performs on par with existing supervised models. Though production applications of our model are limited by ethical considerations, the model's competitive performance points to the great potential of using LLMs for tasks that otherwise require skill-intensive annotation.

What You See is What You Read? Improving Text-Image Alignment Evaluation (2023; Link)

Abstract:
Automatically determining whether a text and a corresponding image are semantically aligned is a significant challenge for vision-language models, with applications in generative text-to-image and image-to-text tasks. In this work, we study methods for automatic image-text alignment evaluation. We first introduce a comprehensive evaluation set spanning multiple datasets from both text-to-image and image-to-text generation tasks, with human judgements for whether a given text-image pair is semantically aligned. We then describe two automatic methods to determine alignment: the first involving a pipeline based on question generation and visual question answering models, and the second employing an end-to-end classification approach based on synthetic data generation. Both methods surpass prior approaches in various text-image alignment tasks, with our analysis showing significant improvements in challenging cases that involve complex composition or unnatural images. Finally, we demonstrate how our approaches can localize specific misalignments between an image and a given text, and how they can be used to automatically re-rank candidates in text-to-image generation.

All You May Need for VQA are Image Captions (NAACL 2022; Link)

Abstract:
Visual Question Answering (VQA) has benefited from increasingly sophisticated models, but has not enjoyed the same level of engagement in terms of data creation. In this paper, we propose a method that automatically derives VQA examples at volume, by leveraging the abundance of existing image-caption annotations combined with neural models for textual question generation. We show that the resulting data is of high-quality. VQA models trained on our data improve state-of-the-art zero-shot accuracy by double digits and achieve a level of robustness that lacks in the same model trained on human-annotated VQA data.

Conciseness: An Overlooked Language Task (EMNLP 2022; Link)

Abstract:
We report on novel investigations into training models that make sentences concise. We define the task and show that it is different from related tasks such as summarization and simplification. For evaluation, we release two test sets, consisting of 2000 sentences each,

that were annotated by two and five raters, respectively. We demonstrate that conciseness is a difficult task for which zero-shot setups with giant neural language models often do not perform well. Given the limitations of these approaches, we propose a synthetic data generation method based on round-trip translations. Using this data to either train Transformers from scratch or fine-tune T5 models yields our strongest baselines that can be further improved by fine-tuning on an artificial conciseness dataset that we derived from multi-annotator machine translation test sets.

Deduplicating Training Data Makes Language Models Better (2022; Link)

Abstract:
As large language models scale up, researchers and engineers have chosen to use larger datasets of loosely-filtered internet text instead of curated texts. We find that existing NLP datasets are highly repetitive and contain duplicated examples. For example, there is an example in the training dataset C4 that has over 200,000 near duplicates. As a whole, we find that 1.68% of the C4 are near-duplicates. Worse, we find a 1% overlap between the training and testing sets in these datasets. Duplicate examples in training data inappropriately biases the distribution of rare/common sequences. Models trained with non-deduplicated datasets are more likely to generate ``memorized" examples. Additionally, if those models are used for downstream applications, such as scoring likelihoods of given sequences, we find that models trained on non-deduplicated and deduplicated datasets have a difference in accuracy of on average TODO.

Description-Driven Task-Oriented Dialog Modeling (ACL 2022; Link)

Abstract:
Task-oriented dialogue (TOD) systems are required to identify key information from conversations for the completion of given tasks. Such information is conventionally specified in terms of intents and slots contained in task-specific ontology or schemata. Since these schemata are designed by system developers, the naming convention for slots and intents is not uniform across tasks, and may not convey their semantics effectively. This can lead to models memorizing arbitrary patterns in data, resulting in suboptimal performance and generalization. In this paper, we propose that schemata should be modified by replacing names or notations entirely with natural language descriptions. We show that a language description-driven system exhibits better understanding of task specifications, higher performance on state tracking, improved data efficiency, and effective zero-shot transfer to unseen tasks. Following this paradigm, we present a simple yet effective Description-Driven Dialog State Tracking (D3ST) model, which relies purely on schema descriptions and an "index-picking" mechanism. We demonstrate the superiority in quality, data efficiency and robustness of our approach as measured on the MultiWOZ (Budzianowski et al.,2018), SGD (Rastogi et al., 2020), and the recent SGD-X (Lee et al., 2021) benchmarks.

Detecting Unintended Memorization in Language-Model-Fused ASR (Proc. Interspeech 2022; Link)

Abstract:
End-to-end (E2E) models are often being accompanied by language models (LMs) via shallow fusion for boosting their overall quality as well as recognition of rare words. At the same time, several prior works show that LMs are susceptible to unintentionally memorizing rare or unique sequences in the training data. In this work, we design a framework for detecting memorization of random textual sequences (which we call canaries) in the LM training data when one has only black-box (query) access to LM-fused speech recognizer, as opposed to direct access to the LM. On a production-grade Conformer RNN-T E2E model fused with a Transformer LM, we show that detecting memorization of singly-occurring canaries from the LM training data of 300M examples is possible. Motivated to protect privacy, we also show that such memorization gets significantly reduced by per-example gradient-clipped LM training without compromising overall quality.

Emergent abilities of large language models (TMLR 2022; Link)

Abstract:
Scaling up language models has been shown to predictably confer a range of benefits such as improved performance and sample efficiency. This paper discusses an unpredictable phenomenon that we call emergent abilities of large language models. Such emergent abilities have close to random performance until evaluated on a model of sufficiently large scale, and hence their emergence cannot be predicted by extrapolating a scaling law based on small-scale models. The emergence of such abilities suggests that additional scaling could further expand the range of tasks that language models can perform. We discuss the implications of these phenomena and suggest directions for future research.

Explainable AI for Practitioners: Designing and implementing explainable ML solutions (O'Reilly Media 2022; Link)

Abstract:
Explainable AI refers to methods and techniques in artificial intelligence (AI) that allow the results of the model to be explained in terms that are understandable by human experts. Explainability is one of the key components of what is now referred to as Responsible AI alongside ML fairness, security and privacy. A successful XAI system aims to increase trust and transparency for complex ML models in a way that benefits model developers, stakeholders, and users.

This book is a collection of some of the most effective and commonly used techniques for explaining why an ML model makes the predictions it does. We discuss the many aspects of Explainable AI including the challenges, metrics for success, and use case studies to guide best practices. Ultimately the goal of this book is to bridge the gap between the vast amount of work that has been done in Explainable AI and provide a quick reference for practitioners that aim to implement XAI into their ML development workflow.

Exploring Length Generalization in Large Language Models (NeurIPS Oral 2022; Link)

Abstract:
The ability to extrapolate from short problem instances to longer ones is an important form of out-of-distribution generalization in reasoning tasks, and is crucial when learning from datasets where longer problem instances are rare. These include theorem proving, solving quantitative mathematics problems, and reading/summarizing novels. In this paper, we run careful empirical studies exploring the length generalization capabilities of transformer-based language models. We first establish that naively finetuning transformers on length generalization tasks shows significant generalization deficiencies independent of model scale. We then show that combining pretrained large language models' in-context learning abilities with scratchpad prompting (asking the model to output solution steps before producing an answer) results in a dramatic improvement in length generalization. We run careful failure analyses on each of the learning modalities and identify common sources of mistakes that highlight opportunities in equipping language models with the ability to generalize to longer problems.

Finetuned Language Models are Zero-Shot Learners (ICLR 2022; Link)

Abstract:
This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that instruction tuning---finetuning language models on a collection of tasks described via instructions---substantially boosts zero-shot performance on unseen tasks.

We take a 137B parameter pretrained language model and instruction-tune it on over 60 NLP tasks verbalized via natural language instruction templates. We evaluate this instruction-tuned model, which we call FLAN, on unseen task types. FLAN substantially improves the performance of its unmodified counterpart and surpasses zero-shot 175B GPT-3 on 20 of 25 tasks that we evaluate. FLAN even outperforms few-shot GPT-3 by a large margin on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, and StoryCloze. Ablation

studies reveal that number of tasks and model scale are key components to the success of instruction tuning.

FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction (ACL 2022; Link)

Abstract:
Sequence modeling has demonstrated state-of-the-art performance on natural language and document understanding tasks. However, it is challenging to correctly serialize tokens in form-like documents in practice due to their variety of layout patterns. We propose FormNet, a structure-aware sequence model to mitigate the suboptimal serialization of forms. First, we design Rich Attention that leverages the spatial relationship between tokens in a form for more precise attention score calculation. Second, we construct Super-Tokens for each word by embedding representations from their neighboring tokens through graph convolutions. FormNet therefore explicitly recovers local syntactic information that may have been lost during serialization. In experiments, FormNet outperforms existing methods with a more compact model size and less pre-training data, establishing new state-of-the-art performance on CORD, FUNSD and Payment benchmarks.

HyperPrompt: Prompt-based Task-Conditioning of Transformers (ICML 2022; Link)

Abstract:
Prompt-tuning is becoming a new paradigm for finetuning pre-trained language models in a parameter-efficient way. Here, we explore the use of HyperNetworks to generate prompts. We propose a novel architecture of HyperPrompt: prompt-based task-conditioned parameterization of self-attention in Transformers. We show that HyperPrompt is very competitive against strong multi-task learning baselines with only 1% of additional task-conditioning parameters. The prompts are end-to-end learnable via generation by a HyperNetwork. The additional parameters scale sub-linearly with the number of downstream tasks, which makes it very parameter efficient for multi-task learning. Hyper-Prompt allows the network to learn task-specific feature maps where the prompts serve as task global memories. Information sharing is enabled among tasks through the HyperNetwork to alleviate task conflicts during co-training. Through extensive empirical experiments, we demonstrate that HyperPrompt can achieve superior performances over strong T5 multi-task learning base-lines and parameter-efficient adapter variants including Prompt-Tuning on Natural Language Understanding benchmarks of GLUE and Super-GLUE across all the model sizes explored.

InnerMonologue: Embodied Reasoning through Planning with Language Models (Conference on Robot Learning 2022; Link)

Abstract:
Recent works have shown the capabilities of large language models to perform tasks requiring reasoning and to be applied to applications beyond natural language processing, such as planning and interaction for embodied robots.These embodied problems require an agent to understand the repertoire of skills available to a robot and the order in which they should be applied. They also require an agent to understand and ground itself within the environment.

In this work we investigate to what extent LLMs can reason over sources of feedback provided through natural language. We propose an inner monologue as a way for an LLM to think through this process and plan. We investigate a variety of sources of feedback, such as success detectors and object detectors, as well as human interaction. The proposed method is validated in a simulation domain and on real robotic. We show that Innerlogue can successfully replan around failures, and generate new plans to accommodate human intent.

LaMDA: Language Models for Dialog Applications (2022; Link)

Abstract:
We present LaMDA: Language Models for Dialog Applications. LaMDA is a family of Transformer-based neural language models specialized for dialog, which have up to 137B parameters and arepre-trained on 1.56T words of public dialog data and web text. While model scaling alone canimprove quality, it shows less improvements on safety and factual grounding. We demonstrate thatfine-tuning with annotated data and enabling the model to consult external knowledge sources canlead to significant improvements towards the two key challenges of safety and factual grounding.The first challenge, safety, involves ensuring that the model's responses are consistent with a set ofhuman values, such as preventing harmful suggestions and unfair bias. We quantify safety using ametric based on an illustrative set of values, and we find that filtering candidate responses using aLaMDA classifier fine-tuned with a small amount of crowdworker-annotated data offers a promisingapproach to improving model safety. The second challenge, factual grounding, involves enabling themodel to consult external knowledge sources, such as an information retrieval system, a languagetranslator, and a calculator. We quantify factuality using a groundedness metric, and we find that ourapproach enables the model to generate responses grounded in known sources, rather than responsesthat merely sound plausible. Finally, we explore the use of LaMDA in the domains of education andcontent recommendations, and analyze their helpfulness and role consistency.

Last Words: Boring Problems are Sometimes the Most Interesting (CL 2022; Link)

Abstract:

In a recent position paper, Turing Award Winners Yoshua Bengio, Geoffrey Hinton and Yann LeCun, make the case that symbolic methods are not needed in AI and that, while there are still many issues to be resolved, AI will be solved using purely neural methods. In this piece I issue a challenge: demonstrate that a purely neural approach to the problem of text normalization is possible. Various groups have tried, but so far nobody has eliminated the problem of unrecoverable errors, errors where, due to insufficient training data or faulty generalization, the system substitutes some other reading for the correct one. Solutions have been proposed that involve a marriage of traditional finite-state methods with neural models, but thus far nobody has shown that the problem can be solved using neural methods alone. Though text normalization is hardly an "exciting" problem, I argue that until one can solve "boring"'problems like that using purely AI methods, one cannot claim that AI is a success.

Long Range Language Modeling via Gated State Spaces (2022; Link)

Abstract:
State space models have shown to be effective for modeling long range dependencies, specifically on sequence classification tasks. In this paper we focus on autoregressive sequence modeling over natural language, Github code and ArXiv mathematics articles. Based on a few recent developments around effectiveness of gated activation functions, we propose a new layer, named Gated State Space (GSS) layer. We show that GSS trains significantly faster than the diagonal version of S4 (i.e. DSS) on TPUs, is simple to implement and fairly competitive with several well-tuned Transformer-based baselines. Finally, we show that interleaving traditional Transformer blocks with GSS improves performance even further.

LongT5: Efficient Text-To-Text Transformer for Long Sequences (NAACL 2022; Link)

Abstract:
Recent work has shown that either (1) increasing the input length or (2) increasing model size can improve the performance of Transformer-based neural models. In this paper, we present a new model, called LongT5, with which we explore the effects of scaling both the input length and model size at the same time. Specifically, we integrated attention ideas from long-input transformers (ETC), and adopted pre-training strategies from summarization

pre-training (PEGASUS) into the scalable T5 architecture. The result is a new attention mechanism we call Transient Global (TGlobal), which mimics ETC's local/global attention mechanism, but without requiring additional side-inputs. We are able to achieve state-of-the-art results on several summarization tasks and outperform the original T5 models on question answering tasks.

Mind's Eye: Grounded Language Model Reasoning through Simulation (ICLR 2023; Link)

Abstract:
Successful and effective communication between humans and AI relies on a shared experience of the world. By training solely on written text, current language models (LMs) miss the grounded experience of humans in the real-world—their failure to relate language to the physical world causes knowledge to be misrepresented and obvious mistakes in their reasoning. We present Mind's Eye, a paradigm to ground language model reasoning in the physical world. Given a physical reasoning question, we use a computational physics engine (DeepMind's MuJoCo) to simulate the possible outcomes, and then use the simulation results as part of the input, which enables language models to perform reasoning. Experiments on 39 tasks in a physics alignment benchmark demonstrate that Mind's Eye can improve reasoning ability by a large margin (27.9% zero-shot, and 46.0% few-shot absolute accuracy improvement on average). Smaller language models armed with Mind's Eye can obtain similar performance to models that are 100× larger. Finally, we confirm the robustness of Mind's Eye through ablation studies.

Natural Language Generation (Almost) from Scratch with Truncated Reinforcement Learning (AAAI 2022; Link)

Abstract:
This paper introduces TRUncated ReinForcement Learning for Language (TrufLL), an original approach to train conditional language models from scratch by only using reinforcement learning (RL). As RL methods unsuccessfully scale to large action spaces, we dynamically truncate the vocabulary space using a generic language model. TrufLL thus en ables to train a language agent by solely interacting with its environment without any task-specific prior knowledge; it is only guided with a task-agnostic language model. Interestingly, this approach avoids the dependency to labelled datasets and inherently reduces pre-trained policy flaws such as language or exposure biases. We evaluate TrufLL on two visual question generation tasks, for which we report promising results over performance and language metrics. To our knowledge, it is the first approach that successfully learns a language generation policy (almost) from scratch

PaLM: Scaling Language Modeling with Pathways (2022; Link)

Abstract:
Large language models have been shown to achieve remarkable performance across a variety of natural language tasks using few-shot learning, which drastically reduces the number of task-specific training examples needed to adapt the model to a particular application. To further our understanding of the impact of scale on few-shot learning, we trained a 540-billion parameter, densely activated, Transformer language model, which we call Pathways Language Model PaLM. We trained PaLM on 6144 TPU v4 chips using Pathways, a new ML system which enables highly efficient training across multiple TPU Pods. We demonstrate continued benefits of scaling by achieving state-of-the-art few-shot learning results on hundreds of language understanding and generation benchmarks. On a number of these tasks, PaLM 540B achieves breakthrough performance, outperforming the finetuned state-of-the-art on a suite of multi-step reasoning tasks, and outperforming average human performance on the recently released BIG-bench benchmark. A significant number of BIG-bench tasks showed discontinuous improvements from model scale, meaning that performance steeply increased as we scaled to our largest model. PaLM also has strong capabilities in multilingual tasks and source code generation, which we demonstrate on a wide array of benchmarks. We additionally provide a comprehensive analysis on bias and toxicity, and study the extent of training data memorization with respect to model scale. Finally, we discuss the ethical considerations related to large language models and discuss potential mitigation strategies.

Pseudo label is better than human label (INTERSPEECH 2022; Link)

Abstract:
Human labeling is expensive. Labeling is the most painful step for ML production. It's widely believed that data is the new gold and big tech companies have an unfair advantage. Is it true that unlimited data unlimits model performance? In this study, we show 1k hrs human labeled data is enough for the best ASR model. The model trained with 1k hrs human labels and 26k hrs pseudo labels has better WERs than the model with 27k hrs human labels. Pseudo label training improves WERs of the production model by a significant margin; 5.9 to 5.1 on voice search. It means pseudo label quality is better than human label. To have quality pseudo labels, we utilized recent self/semi-supervised learning for a large ASR model.

Query Refinement Prompts for Closed-Book Long-Form Question Answering (2022; Link)

Abstract:

Large language models (LLMs) have been shown to perform well in answering questions and in producing long-form texts such as stories and explanations, both in few-shot closed-book settings. While the former can be validated using well-known evaluation metrics, the latter is difficult to evaluate. To this end, we investigate the ability of LLMs to do both tasks at once -- to do question answering that requires long-form answers. Such questions tend to be multifaceted, i.e., they may have ambiguities and/or require information from multiple sources. To this end, we define query refinement prompts that encourage LLMs to explicitly express the multifacetedness in questions and generate long-form answers covering multiple facets of the question. Our experiments on two long-form question answering datasets, ASQA and AQuAMuSe, show that using our prompts allows us to outperform fully finetuned models in the closed book setting, as well as achieve results comparable to retrieve-then-generate open-book models.

SKILL: Structured Knowledge Infusion for Large Language Models
(NAACL 2022; Link)

Abstract:
Large language models (LLMs) have demonstrated human-level performance on vast spectrum of natural language tasks. However, whether they could efficiently memorize or learn from an abstract and structured corpus, like knowledge graph, is largely unexplored. In this work, we propose a method to infuse structure knowledge in LLM, by directly training T5 models on factual triples of knowledge graphs. By evaluating on closed-book QA tasks, we show that models pre-trained with our knowledge-infusing method outperform the T5 baselines, and performs competitively with the models pre-trained on natural language sentences that contain the same knowledge. The proposed method has an advantage that no alignment between the knowledge graph and text corpus is required to curate the training data. This make our method adaptable to industrial scale knowledge graph.

Solving Quantitative Reasoning Problems with Language Models
(NeurIPS 2022; Link)

Abstract:
Language models have achieved remarkable performance on a wide range of tasks that require natural language understanding. Nevertheless, state-of-the-art models have generally struggled with tasks that require quantitative reasoning, such as solving mathematics, science, and engineering problems at the college level. To help close this gap, we introduce Minerva, a large language model pretrained on general natural language data and further trained on technical content. The model achieves state-of-the-art performance on technical benchmarks without the use of external tools. We also evaluate our model on over two hundred undergraduate-level problems in physics, biology, chemistry, economics, and other sciences that require quantitative reasoning, and find that the model can correctly answer nearly a third of them.

Source-summary Entity Aggregation in Abstractive Summarization (COLING 2022; Link)

Abstract:
In a discourse, specific entities that are mentioned can later be referred to by a more general description. For example, 'Celine Dion' and 'Justin Bieber' can be referred to by 'Canadian singers' or 'celebrities'. In this work, we study this phenomenon in the context of summarization, where entities drawn from a source text are generalized in the summary. We call such instances 'source-summary entity aggregations'. We categorize and study several types of source-summary entity aggregations in the CNN/Dailymail corpus, showing that they are reasonably frequent. We experimentally analyze the capabilities of three state-of-the-art summarization systems for generating such aggregations within summaries. We also explore how they can be encouraged to generate more aggregations. Our results show that there is significant room for improvement in generating semantically correct and appropriate aggregations.

Sparsely Activated Language Models are Efficient In-Context Learners (2022; Link)

Abstract:
Scaling language models with more data, compute and parameters has driven significant progress in natural language processing. For example, thanks to scaling, GPT-3 was able to achieve strong performance on few-shot learning. However, training these large dense models require significant amounts of computing resources. In this paper, we develop a family of sparsely activated mixture-of-expert language models named \glam (\textbf{G}eneralist \textbf{La}nguage \textbf{M}odel), which can have many more parameters but require significant less training cost than dense models. The largest \glam has 1.2 trillion parameters, which is approximately 7x larger than GPT-3 but can be trained more efficiently. With only 1/3 of energy consumption to train GPT-3, \glam achieves better overall performance on 29 zero-shot and one-shot NLP tasks. For example, \glam gets 75.0\% one-shot exact match accuracy on the TriviaQA test server, a significant improvement over 68.0\% obtained by GPT-3.

Stretching Sentence-pair NLI Models to Reason over Long Documents and Clusters (EMNLP 2022; Link)

Abstract:

Natural Language Inference (NLI) has been extensively studied by the NLP community as a framework for estimating the semantic relation between sentence pairs. While early work identified certain biases in NLI models, recent advancements in modeling and datasets demonstrated promising performance. In this work, we further explore the direct zero-shot applicability of NLI models to real applications, beyond the sentence-pair setting they were trained on. First, we analyze the robustness of these models to longer and out-of-domain inputs. Then, we develop new aggregation methods to allow operating over full documents, reaching state-of-the-art performance on the ContractNLI dataset. Interestingly, we find NLI scores to provide strong retrieval signals, leading to more relevant evidence extractions compared to common similarity-based methods. Finally, we go further and investigate whole document clusters to identify both discrepancies and consensus among sources. In a test case, we find real inconsistencies between Wikipedia pages in different languages about the same topic.

Table-To-Text generation and pre-training with TabT5 (EMNLP 2022; Link)

Abstract:
Encoder-only transformer models have been successfully applied to different table understanding tasks, as in TAPAS (Herzig et al., 2020). A major limitation of these architectures is that they are constrained to classification-like tasks such as cell selection or entailment detection. We present TABT5, an encoder-decoder model that generates natural language text based on tables and textual inputs. TABT5, overcomes the encoder-only limitation by incorporating a decoder component and leverages the input structure with table specific embeddings as well as pre-training. TABT5 achieves new state-of-the-art results on several domains, including spreadsheet formula prediction (15% increase in sequence accuracy), question answering (10% increase in sequence accuracy) and data-to-text generation (2% increas in BLEU).

TableFormer: Robust Transformer Modeling for Table-Text Encoding (ACL 2022; Link)

Abstract:
Understanding tables is an important aspect of natural language understanding. Existing models for table understanding require linearization of table contents in certain levels, where row or column orders are encoded as unwanted biases. Such spurious biases make the model vulnerable to row and column order perturbations. Also, prior work did not explicitly and thoroughly model structural biases, hindering the table-text modeling ability. In this work, we propose a robust table-text encoding architecture TableFormer, where tabular structural biases are incorporated completely through learnable attention biases. TableFormer is invariant to row and column orders, and could understand tables better due to its tabular inductive biases. Experiments showed that TableFormer outperforms strong baselines in all

settings on SQA, WTQ and TabFact table reasoning datasets, and achieves state-of-the-art performance on SQA, especially when facing answer-invariant row and column perturbations (6% improvement over the best baseline), because previous SOTA models' performance drops by 4% - 6% when facing such perturbations while TableFormer is not affected.

Transformer-based Models of Text Normalization for Speech Applications (2022; Link)

Abstract:
Text normalization, or the process of transforming text into a consistent, canonical form, is crucial for speech applications such as text-to-speech synthesis (TTS). In TTS, the system must decide whether to verbalize "1995" as "nineteen ninety five" in "born in 1995" or as "one thousand nine hundred ninety five" in "page 1995". We present an experimental comparison of various Transformer-based sequence-to-sequence (seq2seq) models of text normalization for speech and evaluate them on a variety of datasets of written text aligned to its normalized spoken form. These models include variants of the 2-stage RNN-based tagging/seq2seq architecture introduced by Zhang et al (2019) where we replace the RNN with a Transformer in one or more stages. We evaluate the performance when initializing the encoder with a pre-trained BERT model. We compare these model variants with a vanilla Transformer that outputs string representations of edit sequences. Of our approaches, using Transformers for sentence context encoding within the 2-stage model proved most effective, with the fine-tuned BERT model yielding the best performance.

A Survey on Data Augmentation Approaches for NLP (ACL 2021; Link)

Abstract:
Data augmentation has recently seen increased interest in NLP due to more work in low-resource domains, new tasks, and the popularity of large-scale neural networks that require large amounts of training data. Despite this recent upsurge, this area is still relatively underexplored, perhaps due to the challenges posed by the discrete nature of language data. In this paper, we present a comprehensive and unifying survey of data augmentation for NLP by summarizing the literature in a structured manner. We first introduce and motivate data augmentation for NLP, and then discuss major methodologically representative approaches. Next, we highlight techniques that are used for popular NLP applications and tasks. We conclude by outlining current challenges and directions for future research. Overall, our paper aims to clarify the landscape of existing literature in data augmentation for NLP and motivate additional work in this area.

AgreeSum: Agreement-Oriented Multi-Document Summarization (ACL 2021; Link)

Abstract:
We aim to renew interest in a particular multi-document summarization (MDS) task which we call AgreeSum: agreement-oriented multi-document summarization. Given a cluster of articles, the goal is to provide abstractive summaries that represent information common and faithful to all input articles. Given the lack of existing datasets, we create a dataset for AgreeSum, and provide annotations on article-summary entailment relations for a subset of the clusters in the dataset. We aim to create strong baselines for the task by applying the top-performing pretrained single-document summarization model PEGASUS onto AgreeSum, leveraging both annotated clusters by supervised losses, and unannotated clusters by T5-based entailment-related and language-related losses. Compared to other baselines, both automatic evaluation and human evaluation show better article-summary and cluster-summary entailment in generated summaries. On a separate note, we hope that our article-summary entailment annotations contribute to the community's effort in improving abstractive summarization faithfulness.

CoMSum: Dataset and Neural Model for Contextual Multi-Document Summarization (ICDAR 2021; Link)

Abstract:
Summarization is the task of compressing source document(s) into coherent and succinct passages. Query-based (contextual) multi-document summarization (qMDS) is a variant that targets summaries to specific informational needs with queries providing additional contexts. Progress in qMDS has been hampered by limited availability of corresponding types of datasets. In this work, we make two contributions. First, we develop an automatic approach for creating both extractive and abstractive qMDS examples from existing language resources. We use this approach to create \qmds, a qMDS dataset for public use. Secondly, to validate the utility of \qmds, we propose a neural model for extractive summarization that exploits the hierarchical nature of the input from multiple documents. It also infuses queries into the modeling to extract query-specific summaries. The experimental results show that modeling the queries and the multiple documents hierarchically improve the performance of qMDS on this datasets. This is consitent with our intuition and supports using \qmds for developing learning methods for qMDS.

Context-Based Quotation Recommendation (ICWSM 2021; Link)

Abstract:

While composing a new document, anything from a news article to an email or essay, authors often utilize direct quotes from a variety of sources. Although an author may know what point they would like to make, selecting an appropriate quote for the specific context may be time-consuming and difficult. We therefore propose a novel context-aware quote recommendation system which utilizes the content an author has already written to generate a ranked list of quotable paragraphs and spans of tokens from a given source document.

We approach quote recommendation as a variant of open-domain question answering and adapt the state-of-the-art BERT-based methods from open-QA to our task. We conduct experiments on a collection of speech transcripts and associated news articles, evaluating models' paragraph ranking and span prediction performances. Our experiments confirm the strong performance of BERT-based methods on this task, which outperform bag-of-words and neural ranking baselines by more than 30% relative across all ranking metrics. Qualitative analyses show the difficulty of the paragraph and span recommendation tasks and confirm the quotability of the best BERT model's predictions, even if they are not the true selected quotes from the original news articles.

COSMOS: Catching out-of-Context Misinformation with Self-Supervised Learning (2021; Link)

Abstract:
Despite the recent attention to DeepFakes, one of the most prevalent ways to mislead audiences on social media is the use of unaltered images in a new but false context. To address these challenges and support fact-checkers, we propose a new method that automatically detects out-of-context image and text pairs. Our key insight is to leverage the grounding of image with text to distinguish out-of-context scenarios that cannot be disambiguated with language alone. We propose a self-supervised training strategy where we only need a set of captioned images. At train time, our method learns to selectively align individual objects in an image with textual claims, without explicit supervision. At test time, we check if both captions correspond to the same object(s) in the image but are semantically different, which allows us to make fairly accurate out-of-context predictions. Our method achieves 85% out-of-context detection accuracy. To facilitate benchmarking of this task, we create a large-scale dataset of 200K images with 450K textual captions from a variety of news websites, blogs, and social media posts. The dataset and source code is publicly available at this https URL.

Data-Efficient Information Extraction from Form-Like Documents (KDD 2021; Link)

Abstract:
Automating information extraction from form-like documents at scale is a pressing need due to its potential impact on automating business workflows across many industries like

financial services, insurance, and healthcare. The key challenge is that form-like documents in these business workflows can be laid out in virtually infinitely many ways; hence, a good solution to this problem should generalize to documents with unseen layouts and languages. A solution to this problem requires a holistic understanding of both the textual segments and the visual cues within a document, which is non-trivial. While the natural language processing and computer vision communities are starting to tackle this problem, there has not been much focus on (1) data-efficiency, and (2) ability to generalize across different document types and languages.

In this paper, we show that when we have only a small number of labeled documents for training (~50), a straightforward transfer learning approach from a considerably structurally-different larger labeled corpus yields up to a 27 F1 point improvement over simply training on the small corpus in the target domain. We improve on this with a simple multi-domain transfer learning approach, that is currently in production use, and show that this yields up to a further 8 F1 point improvement. We make the case that data efficiency is critical to enable information extraction systems to scale to handle hundreds of different document-types, and learning good representations is critical to accomplishing this.

MURAL: Multimodal, Multitask Retrieval Across Languages (EMNLP 2021; Link)

Abstract:
Both image-caption pairs and translation pairs provide the means to learn deep representations of and connections between languages. We use both types of pairs in MURAL (MUltimodal, MUltitask Representations Across Languages), a dual encoder that solves two tasks: 1) image-text matching and 2) translation pair matching. By incorporating billions of translation pairs, MURAL extends ALIGN \cite{jia2021scaling}--a state-of-the-art dual encoder learned from 1.8 billion noisy image-text pairs. When using the same encoders, MURAL's performance matches or exceeds ALIGN's cross-modal retrieval performance on well-resourced languages across several datasets; more importantly, it considerably improves performance on under-resourced languages, showing that text-text learning can overcome a paucity of image-caption examples for these languages. On the Wikipedia Image-Text dataset, for example, MURAL improves zero-shot mean recall by 14.4\% on average for eight under-resourced languages and by 6.6\% on average when fine-tuning. Interestingly, we also find that text representations learned from MURAL cluster based on areal linguistics as well, like the Balkan sprachbund, and not just language genealogy.

Open Domain Question Answering over Tables via Dense Retrieval (NAACL 2021; Link)

Abstract:

Recent advances in open-domain QA have led to strong models based on dense retrieval, but only focused on retrieving textual passages. In this work, we tackle open-domain QA over tables for the first time, and show that retrieval can be improved by a retriever designed to handle tabular context. We present an effective pre-training procedure for our retriever and improve retrieval quality with mined hard negatives. As relevant datasets are missing, we extract a subset of NATURAL QUESTIONS (Kwiatkowski et al., 2019) into a Table QA dataset. We find that our retriever improves retrieval results from 72.0 to 81.1 recall@10 and end-to-end QA results from 33.8 to 37.7 exact match, over a BERT based retriever

ReadTwice: Reading Very Large Documents with Memories (NAACL 2021; Link)

Abstract:
Knowledge-intensive tasks such as question answering often require assimilating information from different sections of large inputs, like books or collections of articles. We propose ReadTwice, a simple and effective approach to combine the advantages of existing approaches that modify Transformers to model long-range dependencies. The main idea is to read smaller segments of the text and summarize them into a memory table to be used in a second read of the text. We show that the model outperforms models of comparable size on several QA datasets and sets the state of the art on the challenging NarrativeQA dataset which asks questions about entire books.

Semi-supervised batch active learning via bilevel optimization (ICASSP 2021; Link)

Abstract:
\emph{Active learning} is an effective technique for reducing the labeling cost by improving data efficiency. In this work, we propose a novel \emph{batch acquisition strategy} for active learning in the setting when the model training is performed in a \emph{semi-supervised} manner. We formulate our approach as a \emph{data summarization} problem via \emph{bilevel optimization}, where the queried batch consists of the points that best summarize the unlabeled data pool. We show that our method is highly effective in \emph{keyword detection} tasks in the regime when only \emph{few labeled samples} are available.

What's the best place for an AI conference, Vancouver or : Why completing comparative questions is difficult (AAAI 2021; Link)

Abstract:
Although large neural language models (LMs) like BERT can be finetuned to yield state-of-the-art results on many NLP tasks, it is often unclear what these models actually

learn. Here we study using such LMs to fill in entities in comparative questions, like “Which country is older, India or ___?”—i.e., we study the ability of neural LMs to ask (not answer) reasonable questions. We show that accuracy in this fill-in-the-blank task is well-correlated with human judgements of whether a question is reasonable, and that these models can be trained to achieve nearly human-level performance in completing comparative questions in three different sub-domains. However, analysis shows that what they learn fails to model any sort of broad notion of which entities are semantically comparable or similar—instead the trained models are very domain-specific, and performance is highly correlated with co-occurrences between specific entities observed in the training set. This is true both for models that are pre-trained on general text corpora, as well as models trained on a large corpus of comparison questions. Our study thus reinforces recent results on the difficulty of making claims about a deep model’s world knowledge or linguistic competence based on performance on specific benchmark problems. We make our evaluation datasets publicly available to foster future research.

Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering (ACL 2021; Link)

Abstract:
Many Question-Answering (QA) datasets contain unanswerable questions, but their treatment in QA systems remains primitive. Our analysis of the Natural Questions (Kwiatkowski et al., 2019) dataset reveals that a substantial portion of unanswerable questions (~21%) can be explained based on the presence of unverifiable presuppositions. Through a user preference study, we demonstrate that the oracle behavior of our proposed system—which provides responses based on presupposition failure—is preferred over the oracle behavior of existing QA systems. Then, we present a novel framework for implementing such a system in three steps: presupposition generation, presupposition verification, and explanation generation, reporting progress on each. Finally, we show that a simple modification of adding presuppositions and their verifiability to the input of a competitive end-to-end QA system yields modest gains in QA performance and unanswerability detection, demonstrating the promise of our approach.

"I’d rather just go to bed”: Understanding Indirect Answers (EMNLP 2020; Link)

Abstract:
We revisit a pragmatic inference problem in dialog: Understanding indirect responses to questions. Humans can interpret I'm starving.' in response toHungry?', even without direct cue words such as yes' andno'. In dialog systems, allowing natural responses rather than closed vocabularies would be similarly beneficial. However, today's systems are only as sensitive to these pragmatic moves as their language model allows. We create and release the first large-scale English language corpus `Circa' with 34,268 (polar question, indirect

answer) pairs to enable progress on this task. The data was collected via elaborate crowd-sourcing, and contains utterances with yes/no meaning, as well as uncertain, middle-ground, and conditional responses. We also present BERT-based neural models to predict such categories for a question-answer pair. We find that while transfer learning from entailment works reasonably, performance is not yet sufficient for robust dialog. Our models reach 82-88% accuracy for a 4-class distinction, and 74-85% for 6 classes.

A Machine of Few Words - Interactive Speaker Recognition with Reinforcement Learning (Interspeech 2020; Link)

Abstract:
Speaker recognition is a well known and studied task in the speech processing domain. It has many applications, either for security or speaker adaptation of personal devices. In this paper, we present a new paradigm for automatic speaker recognition that we call Interactive Speaker Recognition (ISR). In this paradigm, in contrast to the standard text-dependent or text-independent schemes, the recognition system aims at incrementally build a representation of the speakers by requesting personalized utterances to be spoken. To do so, we cast the speaker recognition task into a sequential decision making problem that we solve with Reinforcement Learning. Using a standard dataset, we show that our method achieves very good performance while using little speech signal amounts. This method could also be applied as an utterance selection mechanism for building speech synthesis systems.

Common Conversational Community Prototype: Scholarly Conversational Assistant (2020; Link)

Abstract:
This paper discusses the potential for creating academic resources (tools, data, and evaluation approaches) to support research in conversational search, by focusing on realistic information needs and conversational interactions. Specifically, we propose to develop and operate a prototype conversational search system for scholarly activities. This Scholarly Conversational Assistant would serve as a useful tool, a means to create datasets, and a platform for running evaluation challenges by groups across the community. This article results from discussions of a working group at Dagstuhl Seminar 19461 on Conversational Search.

Controlled Hallucinations:Learning to Generate Faithfully from Noisy Data (EMNLP 2020; Link)

Abstract:

Neural text generation (data- or text-to-text) demonstrates remarkable performance when training data is abundant which for many applications is not the case. To collect a large corpus of parallel data, heuristic rules are often used but they inevitably let noise into the data, such as phrases in the output which cannot be explained by the input. Consequently, models pick up on the noise and may hallucinate–generate fluent but unsupported text. Our contribution is a simple but powerful technique to control such hallucinations without dismissing any input and without modifying the model architecture. On the WikiBio corpus (Lebret et al., 2016), a particularly noisy dataset, we demonstrate the efficacy of the technique both in an automatic and in a human evaluation.

Do Language Embeddings capture Scales? (EMNLP 2020; Link)

Abstract:
We show that embedding-based language models capture a significant amount of information about the scalar magnitudes of objects but are short of the capability required for general common-sense reasoning. We identify ambiguity and numeracy as the key factors limiting their performance, and show that a simple reversible transformation of the pre-training corpus can have a significant effect on the results. We identify the best models and metrics to use when doing zero-shot transfer across tasks in this domain.

Dynamic Composition for Conversational Domain Exploration (ACM 2020; Link)

Abstract:
We study conversational domain exploration (CODEX), where the user's goal is to enrich her knowledge of a given domain by conversing with an informative bot. Such conversations should be well grounded in high-quality domain knowledge as well as engaging and open-ended. A CODEX bot should be proactive and introduce relevant information even if not directly asked for by the user. The bot should also appropriately pivot the conversation to undiscovered regions of the domain. To address these dialogue characteristics, we introduce a novel approach termed dynamic composition that decouples candidate content generation from the flexible composition of bot responses. This allows the bot to control the source, correctness and quality of the offered content, while achieving flexibility via a dialogue manager that selects the most appropriate contents in a compositional manner. We implemented a CODEX bot based on dynamic composition and integrated it into the Google Assistant. As an example domain, the bot conversed about the NBA basketball league in a seamless experience, such that users were not aware whether they were conversing with the vanilla system or the one augmented with our CODEX bot. Results are positive and offer insights into what makes for a good conversation. To the best of our knowledge, this is the first real user experiment of open-ended dialogues as part of a commercial assistant system.

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (ICLR 2020; Link)

Abstract:
Masked language modeling (MLM) pre-training methods such as BERT corrupt the input by replacing some tokens with [MASK] and then train a model to reconstruct the original tokens. While they produce good results when transferred to downstream NLP tasks, they generally require large amounts of compute to be effective. As an alternative, we propose a more sample-efficient pre-training task called replaced token detection. Instead of masking the input, our approach corrupts it by replacing some tokens with plausible alternatives sampled from a small generator network. Then, instead of training a model that predicts the original identities of the corrupted tokens, we train a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. Thorough experiments demonstrate this new pre-training task is more efficient than MLM because the task is defined over all input tokens rather than just the small subset that was masked out. As a result, the contextual representations learned by our approach substantially outperform the ones learned by BERT given the same model size, data, and compute. The gains are particularly strong for small models; for example, we train a model on one GPU for 4 days that outperforms GPT (trained using 30x more compute) on the GLUE natural language understanding benchmark. Our approach also works well at scale, where it performs comparably to RoBERTa and XLNet while using less than 1/4 of their compute and outperforms them when using the same amount of compute.

FELIX: Flexible Text-Editing through Tagging and Insertion (EMNLP 2020; Link)

Abstract:
We present FELIX --- a flexible text-editing approach for generation, designed to derive maximum benefit from the ideas of decoding with bi-directional contexts and self-supervised pre-training. In contrast to conventional sequence-to-sequence (seq2seq) models, FELIX is efficient in low-resource settings and fast at inference time, while being capable of modeling flexible input-output transformations. We achieve this by decomposing the text-editing task into two sub-tasks: tagging to decide on the subset of input tokens and their order in the output text and insertion to in-fill the missing tokens in the output not present in the input. The tagging model employs a novel Pointer mechanism, while the insertion model is based on a Masked Language Model. Both of these models are chosen to be non-autoregressive to guarantee faster inference. FELIX performs favourably when compared to recent text-editing methods and strong seq2seq baselines when evaluated on four NLG tasks: Sentence Fusion, Machine Translation Automatic Post-Editing, Summarization, and Text Simplification.

GoEmotions: A Dataset of Fine-Grained Emotions (ACL 2020; Link)

Abstract:
Understanding emotion expressed in language has a wide range of applications, from building empathetic chatbots to detecting harmful online behavior. Advancement in this area can be improved using large-scale datasets with a fine-grained typology, adaptable to multiple downstream tasks. We introduce GoEmotions, the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We demonstrate the high quality of the annotations via Principal Preserved Component Analysis. We conduct transfer learning experiments with existing emotion benchmarks to show that our dataset generalizes well to other domains and different emotion taxonomies. Our BERT-based model achieves an average F1-score of .46 across our proposed taxonomy, leaving much room for improvement.

Human-centric dialog training via offline reinforcement learning (EMNLP 2020; Link)

Abstract:
How can we train a dialog model to produce better conversations by learning from human feedback, without the risk of humans teaching it harmful chat behaviors? We start by hosting models online, and gather human feedback from real-time, open-ended conversations, which we then use to train and improve the models using offline reinforcement learning (RL). We identify implicit conversational cues including language similarity, elicitation of laughter, sentiment, and more, which indicate positive human feedback, and embed these in multiple reward functions. A well-known challenge is that learning an RL policy in an offline setting usually fails due to the lack of ability to explore and the tendency to make over-optimistic estimates of future reward. These problems become even harder when using RL for language models, which can easily have a 20,000 action vocabulary and many possible reward functions. We solve the challenge by developing a novel class of offline RL algorithms. These algorithms use KL-control to penalize divergence from a pre-trained prior language model, and use a new strategy to make the algorithm pessimistic, instead of optimistic, in the face of uncertainty. We test the resulting dialog model with ratings from 80 users in an open-domain setting and find it achieves significant improvements over existing deep offline RL approaches. The novel offline RL method is viable for improving any existing generative dialog model using a static dataset of human feedback.

Learning-to-Rank with BERT in TF-Ranking (2020; Link)

Abstract:
This paper describes a machine learning algorithm for document (re)ranking, in which queries and documents are firstly encoded using BERT [1], and on top of that a

learning-to-rank (LTR) model constructed with TF-Ranking (TFR) [2] is applied to further optimize the ranking performance. This approach is proved to be effective in a public MS MARCO benchmark [3]. Our submissions achieve the best performance for the passage re-ranking task as of March 30, 2020 [4], and the second best performance for the passage full-ranking task as of April 10, 2020 [5], demonstrating the effectiveness of combining ranking losses with BERT representations for document ranking.

Modifying Memories in Transformer Models (ICML 2021; Link)

Abstract:
Large Transformer models have achieved impressive performance in many natural language tasks. In particular, Transformer based language models have been shown to have great capabilities in encoding factual knowledge in their vast amount of parameters. While the tasks of improving the memorization and generalization of Transformers have been widely studied, it is not well known how to make transformers forget specific old facts and memorize new ones. In this paper, we propose a new task of \emph{explicitly modifying specific factual knowledge in Transformer models while ensuring the model performance does not degrade on the unmodified facts}. This task is useful in many scenarios, such as updating stale knowledge, protecting privacy, and eliminating unintended biases stored in the models. We benchmarked several approaches that provide natural baseline performances on this task. This leads to the discovery of key components of a Transformer model that are especially effective for knowledge modifications. The work also provides insights into the role that different training phases (such as pretraining and fine-tuning) play towards memorization and knowledge modification.

On Faithfulness and Factuality in Abstractive Summarization (ACL 2020; Link)

Abstract:
It is well known that the standard likelihood training and approximate decoding objectives in neural text generation models are fundamentally flawed and lead to dull and repetitive responses. We found that these models when tested on abstractive summarization are highly prone to hallucinate content that is either unfaithful to the input document, completely irrelevant or gibberish. We conduct a large scale human evaluation of several state of the art neural abstractive summarization systems including pretrained language models to better understand the types of hallucinations. Furthermore, we study the extent to which the hallucinated content (i) co-occurs with the common linguistic irregularities such as repetition and incoherence, and (ii) can be measured by NLU measures such as textual entailment, question answering and OpenIE-based fact checking.

Pre-Training Transformers as Energy-Based Cloze Models (EMNLP 2020; Link)

Abstract:
We introduce Electric, an energy-based cloze model for representation learning over text. Like BERT, it is a conditional generative model of tokens given their contexts. However, Electric does not use masking or output a full distribution over tokens that could occur in a context. Instead, it assigns a scalar energy score to each input token indicating how likely it is given its context. We train Electric using an algorithm based on noise-contrastive estimation and elucidate how this learning objective is closely related to the recently proposed ELECTRA pre-training method. Electric performs well when transferred to downstream tasks and is particularly effective at producing likelihood scores for text: it reranks speech recognition n-best lists better than language models and much faster than masked language models. Furthermore, it offers a clearer and more principled view of what ELECTRA learns during pre-training.

Representation Learning for Information Extraction from Form-like Documents (ACL 2020; Link)

Abstract:
We propose a novel approach using representation learning for tackling the problem of extracting structured information from form-like document images. We propose an extraction system that uses knowledge of the types of the target fields to generate extraction candidates, and a neural network architecture that learns a dense representation of each candidate based on neighboring words in the document. These learned representations are not only useful in solving the extraction task for unseen document templates from two different domains, but are also interpretable, as we show using loss cases.

Retrieval Augmented Language Model Pre-Training (ICML 2020; Link)

Abstract:
Language model pre-training has been shown to capture a surprising amount of world knowledge, crucial for NLP tasks such as question answering. However, this knowledge is stored implicitly in the parameters of a neural network, requiring ever-larger networks to cover more facts.

To capture knowledge in a more modular and interpretable way, we augment language model pre-training with a latent knowledge retriever, which allows the model to retrieve and

attend over documents from a large corpus such as Wikipedia, used during pre-training, fine-tuning and inference. For the first time, we show how to pre-train such a knowledge retriever in an unsupervised manner, using masked language modeling as the learning signal and backpropagating through a retrieval step that considers millions of documents.

We demonstrate the effectiveness of Retrieval-Augmented Language Model pre-training (REALM) by fine-tuning on the challenging task of Open-domain Question Answering (Open-QA). We compare against state-of-the-art models for both explicit and implicit knowledge storage on three popular Open-QA benchmarks, and find that we outperform all previous methods by a significant margin (4-16% absolute accuracy), while also providing qualitative benefits such as interpretability and modularity.

Social Biases in NLP Models as Barriers for Persons with Disabilities (ACL 2020; Link)

Abstract:
Building equitable and inclusive technologies demands paying attention to how social attitudes towards persons with disabilities are represented within technology. Representations perpetuated by NLP models often inadvertently encode undesirable social biases from the data on which they are trained. In this paper, first we present evidence of such undesirable biases towards mentions of disability in two different NLP models: toxicity prediction and sentiment analysis. Next, we demonstrate that neural embeddings that are critical first steps in most NLP pipelines also contain undesirable biases towards mentions of disabilities. We then expose the topical biases in the social discourse about some disabilities which may explain such biases in the models; for instance, terms related to gun violence, homelessness, and drug addiction are over-represented in discussions about mental illness.

Tapas: Weakly Supervised Table Parsing via Pre-training (ACL 2020; Link)

Abstract:
Answering natural language questions over tables is usually seen as a semantic parsing task. To alleviate the collection cost of full logical forms, one popular approach focuses on weak supervision consisting of denotations instead of logical forms. However, training semantic parsers from weak supervision poses difficulties, and in addition, the generated logical forms are only used as an intermediate step prior to retrieving the denotation. In this paper, we present TAPAS, an approach to question answering over tables without generating logical forms. TAPAS trains from weak supervision, and predicts the denotation by selecting table cells and optionally applying a corresponding aggregation operator to such selection. TAPAS extends BERT's architecture to encode tables as input, initializes from an effective joint pre-training of text segments and tables crawled from Wikipedia, and is trained end-to-end. We experiment with three different semantic parsing datasets, and find that

TAPAS outperforms or rivals semantic parsing models by improving state-of-the-art accuracy on SQA from 55.1 to 67.2 and performing on par with the state-of-the-art on WIKISQL and WIKITQ, but with a simpler model architecture. We additionally find that transfer learning, which is trivial in our setting, from WIKISQL to WIKITQ, yields 48.7 accuracy, 4.2 points above the state-of-the-art.

Template Guided Text Generation for Task-Oriented Dialogue (EMNLP 2020; Link)

Abstract:
Virtual assistants such as Google Assistant, Amazon Alexa, and Apple Siri enable users to interact with a large number of services and APIs on the web using natural language. In this work, we investigate two methods for Natural Language Generation (NLG) using a single domain-independent model across a large number of APIs. First, we propose a schema-guided approach which conditions the generation on a schema describing the API in natural language. Our second method investigates the use of a small number of templates, growing linearly in number of slots, to convey the semantics of the API. To generate utterances for an arbitrary slot combination, a few simple templates are first concatenated to give a semantically correct, but possibly incoherent and ungrammatical utterance. A pre-trained language model is subsequently employed to rewrite it into coherent, natural sounding text. Through automatic metrics and human evaluation, we show that our method improves over strong baselines, is robust to out-of-domain inputs and shows improved sample efficiency.

The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models (EMNLP, ACL 2020; Link)

Abstract:
We present the Language Interpretability Tool (LIT), an open-source platform for visualization and understanding of NLP models. We focus on core questions about model behavior: Why did my model make this prediction? When does it perform poorly? What happens under a controlled change in the input? LIT integrates local explanations, aggregate analysis, and counterfactual generation into a streamlined, browser-based interface to enable rapid exploration and error analysis. We include case studies for a diverse set of workflows, including exploring counterfactuals for sentiment analysis, measuring gender bias in coreference systems, and exploring local behavior in text generation. LIT supports a wide range of models--including classification, seq2seq, and structured prediction--and is highly extensible through a declarative, framework-agnostic API. LIT is under active development, with code and full documentation available at https://github.com/pair-code/lit.

Thieves of Sesame Street: Model Extraction on BERT-based APIs (ICLR 2020; Link)

Abstract:
We study the problem of model extraction in natural language processing, where an adversary with query access to a victim model attempts to reconstruct a local copy of the model. We show that when both the adversary and victim model fine-tune existing pretrained models such as BERT, the adversary does not need to have access to any training data to mount the attack. Indeed, we show that randomly sampled sequences of words, which do not satisfy grammar structures, make effective queries to extract textual models. This is true even for complex tasks such as natural language inference or question answering.

Our attacks can be mounted with a modest query budget of less than $400.The extraction's accuracy can be further improved using a large textual corpus like Wikipedia, or with intuitive heuristics we introduce. Finally, we measure the effectiveness of two potential defense strategies---membership classification and API watermarking. While these defenses mitigate certain adversaries and come at a low overhead because they do not require re-training of the victim model, fully coping with model extraction remains an open problem.

Towards a Human-like Open-Domain Chatbot (2020; Link)

Abstract:
We present Meena, a multi-turn end-to-end open-domain chatbot trained on data mined from public social media and filtered. The model was trained to minimize perplexity of the next token, but we have found evidence that this metric correlates with human judgement of quality. We propose a human judgement metric called Sensibleness and Specificity Average (SSA) which captures key elements of good conversation. Extensive experiments show strong correlation between perplexity and SSA. The fact that Meena scores high on SSA, 72%, on multi-turn evaluation suggests that a human-like chatbot with SSA score of 82% is potentially within reach if we manage to optimize perplexity better.

Understanding tables with intermediate pre-training (EMNLP 2020; Link)

Abstract:
Table entailment, the binary classification task of finding if a sentence is supported or refuted by the content of a table, requires understanding language and table structure as well as numerical and discrete reasoning. While there is extensive work on textual entailment, table entailment is less well studied. We adapt TAPAS (Herzig et al., 2020), a table-based BERT

model, to recognize entailment. Motivated by the benefits of data augmentation, we create a balanced dataset of millions of automatically created training examples which are learned in an intermediate step prior to fine-tuning. This new data is not only useful for table entailment, but also for SQA (Iyyer et al., 2017), a sequential table QA task. To be able to use long examples as input of BERT models, we evaluate table pruning techniques as a pre-processing step to drastically improve the training and prediction efficiency at a moderate drop in accuracy. The different methods set the new state0o-of-the-art on the TabFact (Chen et al., 2020) and SQA datasets.

Version Control of Speaker Recognition Systems (2020; Link)

Abstract:
This paper discusses one of the most challenging practical engineering problems in speaker recognition systems -the version control of models and user profiles. A typical speaker recognition system consists of two stages: the enrollment stage, where a profile is generated from user-provided enrollment audio; and the runtime stage, where the voice identity of the runtime audio is compared against the stored profiles. As technology advances, the speaker recognition system needs to be updated for better performance. However, if the stored user profiles are not updated accordingly, version mismatch will result in meaningless recognition results. In this paper, we describe different version control strategies for different types of speaker recognition systems, according to how they are deployed in the production environment.

What Happens To BERT Embeddings During Fine-tuning? (EMNLP, ACL 2020; Link)

Abstract:
While there has been much recent work studying how linguistic information is encoded in pre-trained sentence representations, comparatively little is understood about how these models change when adapted to solve downstream tasks. Using a suite of analysis techniques (probing classifiers, Representational Similarity Analysis, and model ablations), we investigate how fine-tuning affects the representations of the BERT model. We find that while fine-tuning necessarily makes significant changes, it does not lead to catastrophic forgetting of linguistic phenomena. We instead find that fine-tuning primarily affects the top layers of BERT, but with noteworthy variation across tasks. In particular, dependency parsing reconfigures most of the model, whereas SQuAD and MNLI appear to involve much shallower processing. Finally, we also find that fine-tuning has a weaker effect on representations of out-of-domain sentences, suggesting room for improvement in model generalization.

Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements (EMNLP 2020; Link)

Abstract:
Natural language descriptions of user interface (UI) elements such as alternative text are crucial for accessibility and language-based interaction in general. Yet, these descriptions are constantly missing in mobile UIs. We propose widget captioning, a novel task for automatically generating language descriptions for UI elements from multimodal input including both the image and the structural representations of user interfaces. We collected a largescale dataset for widget captioning with crowdsourcing. Our dataset contains 162,859 language phrases created by human workers for annotating 61,285 UI elements across 21,750 unique UI screens. We thoroughly analyze the dataset, and train and evaluate a set of deep model configurations to investigate how each feature modality as well as the choice of learning strategies impact the quality of predicted captions. The task formulation and the dataset as well as our benchmark models contribute a solid basis for this novel multimodal captioning task that connects language and user interfaces.

Answering Conversational Questions on Structured Data without Logical Forms (ACL 2019; Link)

Abstract:
We present a novel approach to answering sequential questions based on structured objects such as knowledge bases or tables without using a logical form as an intermediate representation. We encode tables as graphs using a graph neural network model based on the Transformer architecture. The answers are then selected from the encoded graph using a pointer network. This model is appropriate for processing conversations around structured data, where the attention mechanism that selects the answer to a question can also be used to resolve conversational references. We demonstrate the validity of this approach with competitive results on the Sequential Question Answering task (SQA) (Iyyer et al., 2017).

BAM! Born-Again Multi-Task Networks for Natural Language Understanding (ACL 2019; Link)

Abstract:
It can be challenging to train multi-task neural networks that outperform or even match their single-task counterparts. To help address this, we propose using knowledge distillation where single-task models teach a multi-task model. We enhance this training with teacher annealing, a novel method that gradually transitions the model from distillation to supervised learning, helping the multi-task model surpass its single-task teachers. We evaluate our approach by multi-task fine-tuning BERT on the GLUE benchmark. Our method consistently improves over standard single-task and multi-task training.

BERT Rediscovers the Classical NLP Pipeline (ACL 2019; Link)

Abstract:
Pre-trained sentence encoders such as ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2018) have rapidly advanced the state-of-theart on many NLP tasks, and have been shown to encode contextual information that can resolve many aspects of language structure. We extend the edge probing suite of Tenney et al. (2019) to explore the computation performed at each layer of the BERT model, and find that tasks derived from the traditional NLP pipeline appear in a natural progression: part-of-speech tags are processed earliest, followed by constituents, dependencies, semantic roles, and coreference. We trace individual examples through the encoder and find that while this order holds on average, the encoder occasionally inverts the order, revising low-level decisions after deciding higher-level contextual relations.

BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions (NAACL 2019; Link)

Abstract:
In this paper we study yes/no questions that are naturally occurring---meaning that they are generated in unprompted and unconstrained settings. We build a reading comprehension dataset, BoolQ, of such questions, and show that they are unexpectedly challenging. They often query for complex, non-factoid information, and require difficult entailment-like inference to solve. We also explore the effectiveness of a range of transfer learning baselines. We find that transferring from entailment data is more effective than transferring from paraphrase or extractive QA data, and that it, surprisingly, continues to be very beneficial even when starting from massive pre-trained language models such as BERT. Our best method trains BERT on MultiNLI and then re-trains it on our train set. It achieves 80.4% accuracy compared to 90% accuracy of human annotators (and 62% majority-baseline), leaving a significant gap for future work.

Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension (EMNLP 2019; Link)

Abstract:
Reading comprehension models have been successfully applied to extractive text answers, but it is unclear how best to generalize these models to abstractive numerical answers. We enable a BERT-based reading comprehension model to perform lightweight numerical reasoning. We augment the model with a predefined set of executable 'programs' which

encompass simple arithmetic as well as extraction. Rather than having to learn to manipulate numbers directly, the model can pick a program and execute it. On the recent Discrete Reasoning Over Passages (DROP) dataset, designed to challenge reading comprehension models, we show a 33% absolute improvement by adding shallow programs. The model can learn to predict new operations when appropriate in a math word problem setting (Roy and Roth, 2015) with very few training examples.

Identifying and Correcting Label Bias in Machine Learning (2019; Link)

Abstract:
Datasets often contain biases which unfairly disadvantage certain groups, and classifiers trained on such datasets can inherit these biases. In this paper, we provide a mathematical formulation of how this bias can arise. We do so by assuming the existence of underlying, unknown, and unbiased labels which are overwritten by an agent who intends to provide accurate labels but may have biases against certain groups. Despite the fact that we only observe the biased labels, we are able to show that the bias may nevertheless be corrected by re-weighting the data points without changing the labels. We show, with theoretical guarantees, that training on the re-weighted dataset corresponds to training on the unobserved but unbiased labels, thus leading to an unbiased machine learning classifier. Our procedure is fast and robust and can be used with virtually any learning algorithm. We evaluate on a number of standard machine learning fairness datasets and a variety of fairness notions, finding that our method outperforms standard approaches in achieving fair classification.

Learning Entity Representations for Few-Shot Reconstruction of Wikipedia Categories (2019; Link)

Abstract:
Language modeling tasks, in which words are predicted on the basis of a local context, have been very effective for learning word embeddings and context dependent representations of phrases. Motivated by the observation that efforts to code world knowledge into machine readable knowledge bases tend to be entity-centric, we investigate the use of a fill-in-the-blank task to learn context independent representations of entities from the contexts in which those entities were mentioned. We show that large scale training of neural models allows us to learn extremely high fidelity entity typing information, which we demonstrate with few-shot reconstruction of Wikipedia categories. Our learning approach is powerful enough to encode specialized topics such as Giro d'Italia cyclists.

Learning to Navigate the Web (ICLR 2019; Link)

Abstract:
Learning in environments with large state and action spaces, and sparse rewards, can hinder a Reinforcement Learning (RL) agent's learning through trial-anderror. For instance, following natural language instructions on the Web (such as booking a flight ticket) leads to RL settings where input vocabulary and number of actionable elements on a page can grow very large. Even though recent approaches improve the success rate on relatively simple environments with the help of human demonstrations to guide the exploration, they still fail in environments where the set of possible instructions can reach millions. We approach the aforementioned problems from a different perspective and propose guided RL approaches that can generate unbounded amount of experience for an agent to learn from. Instead of learning from a complicated instruction with a large vocabulary, we decompose it into multiple sub-instructions and schedule a curriculum in which an agent is tasked with a gradually increasing subset of these relatively easier sub-instructions. In addition, when the expert demonstrations are not available, we propose a novel meta-learning framework that generates new instruction following tasks and trains the agent more effectively. We train DQN, deep reinforcement learning agent, with Q-value function approximated with a novel QWeb neural network architecture on these smaller, synthetic instructions. We evaluate the ability of our agent to generalize to new instructions on World of Bits benchmark, on forms with up to 100 elements, supporting 14 million possible instructions. The QWeb agent outperforms the baseline without using any human demonstration achieving 100% success rate on several difficult environments.

Like a Baby: Visually Situated Neural Language Acquisition (ACL 2019; Link)

Abstract:
We examine the benefits of visual context in training neural language models to perform next-word prediction. A multi-modal neural architecture is introduced that outperform its equivalent trained on language alone with a 2% decrease in perplexity, even when no visual context is available at test. Fine-tuning the embeddings of a pre-trained state-of-the-art bidirectional language model (BERT) in the language modeling framework yields a 3.5% improvement. The advantage for training with visual context when testing without is robust across different languages (English, German and Spanish) and different models (GRU, LSTM, Delta-RNN, as well as those that use BERT embeddings). Thus, language models perform better when they learn like a baby, i.e, in a multi-modal environment. This finding is compatible with the theory of situated cognition: language is inseparable from its physical context.

MeanSum: A Neural Model for Unsupervised Multi-document Abstractive Summarization (ICML 2019; Link)

Abstract:
Abstractive summarization has been studied using neural sequence transduction methods with datasets of large, paired document-summary examples. However, such datasets are rare and the models trained from them do not generalize to other domains. Recently, some progress has been made in learning sequence-to-sequence mappings with only unpaired examples. In our work, we consider the setting where there are only documents and no summaries provided and propose and end-to-end, neural model architecture to perform unsupervised abstractive summarization. Our proposed model consists of an auto-encoder trained so that the mean of the representations of the input documents decodes to a reasonable summary. We consider variants of the proposed architecture and perform an ablation study to show the importance of specific components. We apply our model to the summarization of business and product reviews and show that the generated summaries are fluent, show relevancy in terms of word-overlap, and are representative of the average sentiment with respect to the input documents compared to baselines.

Measuring Domain Portability and Error Propagation in Biomedical QA (BioASQ 2019; Link)

Abstract:
In this work we present Google's submission to the BioASQ 7 biomedical question answering (QA) task (specifically Task 7b, Phase B). The core of our systems are based on BERT QA models, specifically the model of [1]. In this report, and via our submissions, we aimed to investigate two research questions. We start by studying how domain portable are QA systems that have been pre-trained and fine-tuned on general texts, e.g., Wikipedia. We measure this via two submissions. The first is a non-adapted model that uses a public pre-trained BERT model and is fine-tuned on the Natural Questions data set [4]. The second system takes this non-adapted model and fine-tunes it with the BioASQ training data. Next, we study the impact of error propagation in end-to-end retrieval and QA systems. Again we test this via two submissions. The first uses human annotated relevant documents and snippets as input to the model and the second predicted documents and snippets. Our main findings are that domain specific fine-tuning can benefit Biomedical QA. However, the biggest quality bottleneck is at the retrieval stage, where we see large drops in metrics – over 10pts absolute – when using non gold inputs to the QA model.

Model Cards for Model Reporting (2019; Link)

Abstract:

Trained machine learning models are increasingly used to perform high impact tasks such as determining crime recidivism rates and predicting health risks. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts they are not well-suited for, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards (or M-cards) to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic subgroups (e.g., race, geographic location, sex, Fitzpatrick skin tone) and intersectional subgroups (e.g., age and race, or sex and Fitzpatrick skin tone) that are relevant to the intended application domains. Model cards also disclose the context under which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model. To solidify the concept, we provide cards for models trained to detect smiling faces on the CelebA dataset (Liu et al., 2015) and models trained to detect toxicity in the Conversation AI dataset (Dixon et al., 2018). We propose this work as a step towards the responsible democratization of machine learning and related AI technology, providing context around machine learning models and increasing the transparency into how well such models work. We hope this work encourages those releasing trained machine learning models to accompany model releases with similar detailed documentation.

Natural Questions: a Benchmark for Question Answering Research (ACL 2019; Link)

Abstract:
We present the Natural Questions corpus, a question answering dataset. Questions consist of real anonymized, aggregated queries issued to the Google search engine. An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present. The public release consists of 307,373 training examples with single annotations, 7,830 examples with 5-way annotations for development data, and a further 7,842 examples 5-way annotated sequestered as test data. We present experiments validating quality of the data. We also describe analysis of 25-way annotations on 302 examples, giving insights into human variability on the annotation task. We introduce robust metrics for the purposes of evaluating question answering systems; demonstrate high human upper bounds on these metrics; and establish baseline results using competitive methods drawn from related literature.

Parameter Efficient Transfer Learning for NLP (ICML 2019; Link)

Abstract:
Fine-tuning large pretrained models is an effective transfer mechanism in NLP. However, in the presence of many downstream tasks, fine-tuning is parameter inefficient: an entire new model is required for every task. As an alternative, we propose transfer with adapter modules. Adapter modules yield a compact and extensible model; they add only a few trainable parameters per task, and new tasks can be added without revisiting previous ones. The parameters of the original network remain fixed, yielding a high degree of parameter sharing. To demonstrate adapter's effectiveness, we transfer the recently proposed BERT Transformer model to 26 diverse text classification tasks, including the GLUE benchmark. Adapters attain near state-of-the-art performance, whilst adding only a few parameters per task. On GLUE, we attain within 0.8% of the performance of full fine-tuning, adding only 3.6% parameters per task. By contrast, fine-tuning trains 100% of the parameters per task.

Personal VAD: Speaker-Conditioned Voice Activity Detection (2019; Link)

Abstract:
In this paper, we propose "personal VAD", a system to detect the voice activity of a target speaker at the frame level. This system is useful for gating the inputs to a streaming speech recognition system, such that it only triggers for the target user, which helps reduce the computational cost and battery consumption. We achieve this by training a VAD-alike neural network which is conditioned on the target speaker embedding or the speaker verification score. For every frame, personal VAD outputs the scores for three classes: non-speech, target speaker speech, and non-target speaker speech. With our optimal setup, we are able to train a 130KB model that out-performs a baseline system where individually trained standard VAD and speaker recognition network are combined to perform the same task.

Perturbation Sensitivity Analysis to Detect Unintended Model Biases (EMNLP, ACL 2019; Link)

Abstract:
Data-driven statistical Natural Language Processing (NLP) techniques leverage large amounts of language data to build models that can understand language. However, most language data reflect the public discourse at the time the data was produced, and hence NLP models are susceptible to learning incidental associations around named referents at a particular point in time, in addition to general linguistic meaning. An NLP system designed to model notions such as sentiment and toxicity should ideally produce scores that are independent of the identity of such entities mentioned in text and their social associations. For example, in a general purpose sentiment analysis system, a phrase such as I hate Katy Perry should be interpreted as having the same sentiment as I hate Taylor Swift. Based on this idea, we propose a generic evaluation framework, Perturbation Sensitivity Analysis, which detects unintended model biases related to named entities, and requires no new

annotations or corpora. We demonstrate the utility of this analysis by employing it on two different NLP models --- a sentiment model and a toxicity model --- applied on online comments in English language from four different genres.

Reducing Permission Requests in Mobile Apps (IMC 2019; Link)

Abstract:
Users of mobile apps sometimes express discomfort or concerns with what they see as unnecessary or intrusive permission requests by certain apps. However encouraging mobile app developers to request fewer permissions is challenging because there are many reasons why permissions are requested; furthermore, prior work has shown it is hard to disambiguate the purpose of a particular permission with high certainty. In this work we describe a novel, algorithmic mechanism intended to discourage mobile-app developers from asking for unnecessary permissions. Developers are incentivized by an automated alert, or "nudge", shown in the Google Play Console when their apps ask for permissions that are requested by very few functionally-similar apps---in other words, by their competition. Empirically, this incentive is effective, with significant developer response since its deployment. Permissions have been redacted by 59% of apps that were warned, and this attenuation has occurred broadly across both app categories and app popularity levels. Importantly, billions of users' app installs from the Google Play have benefited from these redactions

Semantic Text Matching for Long-Form Documents (WWW Conference 2019; Link)

Abstract:

Semantic text matching is one of the most important research problems in many domains, including, but not limited to, information retrieval, question answering, and recommendation. Among the different types of semantic text matching, long-document-to-long-document text matching has many applications, but has rarely been studied. Most existing approaches for semantic text matching have limited success in this setting, due to their inability to capture and distill the main ideas and topics from long-form text.

In this paper, we propose a novel Siamese multi-depth attention-based hierarchical recurrent neural network (SMASH RNN) that learns the long-form semantics, and enables long-form document based semantic text matching. In addition to word information, SMASH RNN is using the document structure to improve the representation of long-form documents. Specifically, SMASH RNN synthesizes information from different document structure levels, including paragraphs, sentences, and words. An attention-based hierarchical RNN derives a representation for each document structure level. Then, the representations learned from the different levels are aggregated to learn a more comprehensive semantic representation of

the entire document. For semantic text matching, a Siamese structure couples the representations of a pair of documents, and infers a probabilistic score as their similarity.

We conduct an extensive empirical evaluation of SMASH RNN with three practical applications, including email attachment suggestion, related article recommendation, and citation recommendation. Experimental results on public data sets demonstrate that SMASH RNN significantly outperforms competitive baseline methods across various classification and ranking scenarios in the context of semantic matching of long-form documents.

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks (USENIX Security 2019; Link)

Abstract:
This paper describes a testing methodology for quantitatively assessing the risk of \emph{unintended memorization} of rare or unique sequences in generative sequence models---a common type of neural network. Such models are sometimes trained on sensitive data (e.g., the text of users' private messages); our methodology allows deep-learning to choose configurations that minimize memorization during training, thereby benefiting privacy.

In experiments, we show that unintended memorization is a persistent, hard-to-avoid issue that can have serious consequences. Specifically, if not addressed during training, we show that new, efficient procedures can allow extracting unique, secret sequences such as credit card numbers from trained models. We also show that our testing strategy is practical and easy-to-apply, e.g., by describing its use for quantitatively preventing data exposure in a production, commercial neural network---a predictive email-composition assistant trained on millions of users' email messages.

Towards Automatic Concept-based Explanations (NeurIPS 2019; Link)

Abstract:
Interpretability has become an important topic of research as more machine learning (ML) models are deployed and widely used to make important decisions. Most of the current explanation methods provide explanations through feature importance scores, which identify features that are salient for each individual input. However, how to systematically summarize and interpret such per sample feature importance scores itself is challenging. In this work, we propose principles and desiderata for \emph{concept} based explanation, which goes beyond per-sample features to identify higher level human-understandable concepts that apply across the entire dataset. We develop a new algorithm, ACE, to automatically extract visual concepts. Our systematic experiments demonstrate that ACE discovers concepts that are human-meaningful, coherent and salient for the neural network's predictions.

Using Audio Transformations to Improve Comprehension in Voice Question Answering (CLEF 2019; Link)

Abstract:
Many popular form factors of digital assistants—such as Amazon Echo, Apple Homepod, or Google Home—enable the user to hold a conversation with these systems based only on the speech modality. The lack of a screen presents unique challenges. To satisfy the information need of a user, the presentation of the answer needs to be optimized for such voice-only interactions. In this paper, we propose a task of evaluating the usefulness of audio transformations (i.e., prosodic modifications) for voice-only question answering. We introduce a crowdsourcing setup where we evaluate the quality of our proposed modifications along multiple dimensions corresponding to the informativeness, naturalness, and ability of the user to identify key parts of the answer. We offer a set of prosodic modifications that highlight potentially important parts of the answer using various acoustic cues. Our experiments show that some of these modifications lead to better comprehension at the expense of only slightly degraded naturalness of the audio.

A General Approach to Adding Differential Privacy to Iterative Training Procedures (NIPS 2018; Link)

Abstract:
In this work we address the practical challenges of training machine learning models on privacy-sensitive datasets by introducing a modular approach that minimizes changes to training algorithms, provides a variety of configuration strategies for the privacy mechanism, and then isolates and simplifies the critical logic that computes the final privacy guarantees. A key challenge is that training algorithms often require estimating many different quantities (vectors) from the same set of examples --- for example, gradients of different layers in a deep learning architecture, as well as metrics and batch normalization parameters. Each of these may have different properties like dimensionality, magnitude, and tolerance to noise. By extending previous work on the Moments Accountant for the subsampled Gaussian mechanism, we can provide privacy for such heterogeneous sets of vectors, while also structuring the approach to minimize software engineering challenges.

An efficient framework for learning sentence representations (ICLR 2018; Link)

Abstract:

In this work we propose a simple and efficient framework for learning sentence representations from unlabelled data. Drawing inspiration from the distributional hypothesis and recent work on learning sentence representations, we reformulate the problem of predicting the context in which a sentence appears as a classification problem. This allows us to efficiently learn different types of encoding functions, and we show that the model learns high-quality sentence representations. We demonstrate that our sentence representations outperform state-of-the-art unsupervised and supervised representation learning methods on several downstream NLP tasks that involve understanding sentence semantics while achieving an order of magnitude speedup in training time.

AUEB-NLP at BioASQ 6: Document and Snippet Retrieval (2018; Link)

Abstract:
We present AUEB's submissions to the BioASQ 6 document and snippet retrieval tasks (parts of Task 6b, Phase A). Our models use novel extensions to deep learning architectures that operate solely over the text of the query and candidate document/snippets. Our systems scored at the top or near the top for all batches of the challenge, highlighting the effectiveness of deep learning for these tasks.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL 2018; Link)

Abstract:
We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

Eval all, trust a few, do wrong to none: Comparing sentence generation models (2018; Link)

Abstract:
In this paper we study various flavors of variational autoencoders and address the methodological issues with the current neural text generation research and also close some gaps by answering a few natural questions to the studies already published.

Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone (Google AI Blog 2018; Link)

Abstract:
A long-standing goal of human-computer interaction has been to enable people to have a natural conversation with computers, as they would with each other. In recent years, we have witnessed a revolution in the ability of computers to understand and to generate natural speech, especially with the application of deep neural networks (e.g., Google voice search, WaveNet). Still, even with today's state of the art systems, it is often frustrating having to talk to stilted computerized voices that don't understand natural language. In particular, automated phone systems are still struggling to recognize simple words and commands. They don't engage in a conversation flow and force the caller to adjust to the system instead of the system adjusting to the caller.

Today we announce Google Duplex, a new technology for conducting natural conversations to carry out "real world" tasks over the phone. The technology is directed towards completing specific tasks, such as scheduling certain types of appointments. For such tasks, the system makes the conversational experience as natural as possible, allowing people to speak normally, like they would to another person, without having to adapt to a machine.

Identifying Well-formed Natural Language Questions (EMNLP 2019; Link)

Abstract:
Understanding natural language queries is fundamental to many practical NLP systems. Often, such systems comprise of a brittle processing pipeline, that is not robust to "word salad" text ubiquitously issued by users. However, if a query resembles a grammatical and well-formed question, such a pipeline is able to perform more accurate interpretation, thus reducing downstream compounding errors. Hence, identifying whether or not a query is well formed can enhance query understanding. Here, we introduce a new task of identifying a well-formed natural language question. We construct and release a dataset of 25,100 publicly available questions classified into well-formed and non-well-formed categories and report an accuracy of 70.7% on the test set. We also show that our classifier can be used to

improve the performance of neural sequence-to-sequence model for generating questions for reading comprehension.

Illustrative Language Understanding: Large-Scale Visual Grounding with Google Image Search (ACL 2018; Link)

Abstract:
We introduce Picturebook, a large-scale lookup operation to ground language via ‘snapshots’ of our physical world accessed through image search. For each word in a vocabulary, we extract the top-k images from Google image search and feed the images through a convolutional network to extract a word embedding. We introduce a multimodal gating function to fuse our Picturebook embeddings with other word representations. We also introduce Inverse Picturebook, a mechanism to map a Picturebook embedding back into words. We experiment and report results across a wide range of tasks: word similarity, natural language inference, semantic relatedness, sentiment/topic classification, image-sentence ranking and machine translation. We also show that gate activations corresponding to Picturebook embeddings are highly correlated to human judgments of concreteness ratings.

Learning to Attack: Adversarial Transformation Networks (AAAI 2018; Link)

Abstract:
With the rapidly increasing popularity of deep neural networks for image recognition tasks, a parallel interest in generating adversarial examples to attack the trained models has arisen. To date, these approaches have involved either directly computing gradients with respect to the image pixels or directly solving an optimization on the image pixels. We generalize this pursuit in a novel direction: can a separate network be trained to efficiently attack another fully trained network? We demonstrate that it is possible, and that the generated attacks yield startling insights into the weaknesses of the target network. We call such a network an Adversarial Transformation Network (ATN). ATNs transform any input into an adversarial attack on the target network, while being minimally perturbing to the original inputs and the target network’s outputs. Further, we show that ATNs are capable of not only causing the target network to make an error, but can be constructed to explicitly control the type of misclassification made. We demonstrate ATNs on both simple MNIST digit classifiers and state-of-the-art ImageNet classifiers deployed by Google, Inc.: Inception ResNet-v2.

Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning (ICLR 2018; Link)

Abstract:
Deep reinforcement learning algorithms can learn complex behavioral skills, but real-world application of these methods requires a considerable amount of experience to be collected by the agent. In practical settings, such as robotics, this involves repeatedly attempting a task, resetting the environment between each attempt. However, not all tasks are easily or automatically reversible. In practice, this learning process requires considerable human intervention. In this work, we propose an autonomous method for safe and efficient reinforcement learning that simultaneously learns a forward and backward policy, with the backward policy resetting the environment for a subsequent attempt. By learning a value function for the backward policy, we can automatically determine when the forward policy is about to enter a non-reversible state, providing for uncertainty-aware safety aborts. Our experiments illustrate that proper use of the backward policy can greatly reduce the number of manual resets required to learn a task and can reduce the number of unsafe actions that lead to non-reversible states.

State-of-the-art Chinese Word Segmentation with Bi-LSTMs (ACL 2018; Link)

Abstract:
A wide variety of neural-network architectures have been proposed for the task of Chinese word segmentation. Surprisingly, we find that a bidirectional LSTM model, when combined with standard deep learning techniques and best practices, can achieve better accuracy on many of the popular datasets as compared to models based on more complex neuralnetwork architectures. Furthermore, our error analysis shows that out-of-vocabulary words remain challenging for neural-network models, and many of the remaining errors are unlikely to be fixed through architecture changes. Instead, more effort should be made on exploring resources for further improvement.

State-of-the-art Speech Recognition With Sequence-to-Sequence Models (ICASSP 2018; Link)

Abstract:
Attention-based encoder-decoder architectures such as Listen, Attend, and Spell (LAS), subsume the acoustic, pronunciation and language model components of a traditional automatic speech recognition (ASR) system into a single neural network. In our previous work, we have shown that such architectures are comparable to state-of-the-art ASR systems on dictation tasks, but it was not clear if such architectures would be practical for more challenging tasks such as voice search. In this work, we explore a variety of structural

and optimization improvements to our LAS model which significantly improve performance. On the structural side, we show that word piece models can be used instead of graphemes. We introduce a multi-head attention architecture, which offers improvements over the commonly-used single-head attention. On the optimization side, we explore techniques such as synchronous training, scheduled sampling, label smoothing, and minimum word error rate optimization, which are all shown to improve accuracy. We present results with a unidirectional LSTM encoder for streaming recognition. On a 12,500 hour voice search task, we find that the proposed changes improve the WER of the LAS system from 9.2% to 5.6%, while the best conventional system achieve 6.7% WER. We also test both models on a dictation dataset, and our model provide 4.1% WER while the conventional system provides 5% WER.

Text Embeddings Contain Bias. Here's Why That Matters. (Google 2018; Link)

Abstract:
With the public release of embedding models, it’s important to understand the various biases that they contain. Developers who use them should be aware of the biases inherent in the models as well as how biases can manifest in downstream applications that use these models. In this post, we examine a few specific forms of bias and suggest tools for evaluating as well as mitigating bias.

To Trust Or Not To Trust A Classifier (NeurIPS 2018; Link)

Abstract:
Knowing when a classifier's prediction can be trusted is useful in many applications and critical for safely using AI. While the bulk of the effort in machine learning research has been towards improving classifier performance, understanding when a classifier's predictions should and should not be trusted has received far less attention. The standard approach is to use the classifier's discriminant or confidence score; however, we show there exists an alternative that is more effective in many situations. We propose a new score, called the {\it trust score}, which measures the agreement between the classifier and a modified nearest-neighbor classifier on the testing example. We show empirically that high (low) trust scores produce surprisingly high precision at identifying correctly (incorrectly) classified examples, consistently outperforming the classifier's confidence score as well as many other baselines. Further, under some mild distributional assumptions, we show that if the trust score for an example is high (low), the classifier will likely agree (disagree) with the Bayes-optimal classifier. Our guarantees consist of non-asymptotic rates of statistical consistency under various nonparametric settings and build on recent developments in topological data analysis.

Universal Sentence Encoder (EMNLP 2018; Link)

Abstract:
We present models for encoding sentences into embedding vectors that specifically target transfer learning to other NLP tasks. The models are efficient and result in accurate performance on diverse transfer tasks. Two variants of the encoding models allow for trade-offs between accuracy and compute resources. For both variants, we investigate and report the relationship between model complexity, resource consumption, the availability of transfer task training data, and task performance. Comparisons are made with baselines that use word level transfer learning via pretrained word embeddings as well as baselines do not use any transfer learning. We find that transfer learning using sentence embeddings tends to outperform word level transfer. With transfer learning via sentence embeddings, we observe surprisingly good performance with minimal amounts of supervised training data for a transfer task. We obtain encouraging results on Word Embedding Association Tests (WEAT) targeted at detecting model bias. Our pre-trained sentence encoding models are made freely available for download and on TF Hub.

Analyza: Exploring Data with Conversation (ACM 2017; Link)

Abstract:
We describe Analyza, a system that helps lay users explore data. Analyza has been used within two large real world systems. The first is a question-and-answer feature in a spreadsheet product. The second provides convenient access to a revenue/inventory database for a large sales force. Both user bases consist of users who do not necessarily have coding skills, demonstrating Analyza's ability to democratize access to data.

We discuss the key design decisions in implementing this system. For instance, how to mix structured and natural language modalities, how to use conversation to disambiguate and simplify querying, how to rely on the ``semantics'' of the data to compensate for the lack of syntactic structure, and how to efficiently curate the data.

Attention is All You Need (NIPS 2017; Link)

Abstract:
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. 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 less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including

ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Attention-Based Models for Text-Dependent Speaker Verification (2017; Link)

Abstract:
Attention-based models have recently shown great performance on a range of tasks, such as speech recognition, machine translation, and image captioning due to their ability to summarize relevant information that expands through the entire length of an input sequence. In this paper, we analyze the usage of attention mechanisms to the problem of sequence summarization in our end-to-end text-dependent speaker recognition system. We explore different topologies and their variants of the attention layer, and compare different pooling methods on the attention weights. Ultimately, we show that attention-based models can improves the Equal Error Rate (EER) of our speaker verification system by relatively 14% compared to our non-attention LSTM baseline model.

Efficient Natural Language Response Suggestion for Smart Reply (2017; Link)

Abstract:
This paper presents a computationally efficient machine-learned method for natural language response suggestion. Feed-forward neural networks using n-gram embedding features encode messages into vectors which are optimized to give message-response pairs a high dot-product value. An optimized search finds response suggestions. The method is evaluated in a large-scale commercial e-mail application, Inbox by Gmail. Compared to a sequence-to-sequence approach, the new system achieves the same quality at a small fraction of the computational requirements and latency.

Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence LSTM-based autoencoders (Interspeech 2017; Link)

Abstract:
A neural network model that significant improves unit-selection-based Text-To-Speech synthesis is presented. The model employs a sequence-to-sequence LSTM-based autoencoder that compresses the acoustic and linguistic features of each unit to a fixed-size

vector referred to as an embedding. Unit-selection is facilitated by formulating the target cost as an L2 distance in the embedding space. In open-domain speech synthesis the method achieves a 0.2 improvement in the MOS, while for limited-domain it reaches the cap of 4.5 MOS. Furthermore, the new TTS system halves the gap between the previous unit-selection system and WaveNet in terms of quality while retaining low computational cost and latency.

Highway-LSTM and Recurrent Highway Networks for Speech Recognition (Interspeech 2017; Link)

Abstract:
Recently, very deep networks, with as many as hundreds of layers, have shown great success in image classification tasks. One key component that has enabled such deep models is the use of “skip connections”, including either residual or highway connections, to alleviate the vanishing and exploding gradient problems. While these connections have been explored for speech, they have mainly been explored for feed-forward networks. Since recurrent structures, such as LSTMs, have produced state-of-the-art results on many of our Voice Search tasks, the goal of this work is to thoroughly investigate different approaches to adding depth to recurrent structures. Specifically, we experiment with novel Highway-LSTM models with bottlenecks skip connections and show that a 10 layer model can outperform a state-of-the-art 5 layer LSTM model with the same number of parameters by 2% relative WER. In addition, we experiment with Recurrent Highway layers and find these to be on par with Highway-LSTM models, when given sufficient depth.

Human and Machine Hearing: Extracting Meaning from Sound (Cambridge University Press 2017; Link)

Abstract:
Human and Machine Hearing is the first book to comprehensively describe how human hearing works and how to build machines to analyze sounds in the same way that people do. Drawing on over thirty-five years of experience in analyzing hearing and building systems, Richard F. Lyon explains how we can now build machines with close-to-human abilities in speech, music, and other sound-understanding domains. He explains human hearing in terms of engineering concepts, and describes how to incorporate those concepts into machines for a wide range of modern applications. The details of this approach are presented at an accessible level, to bring a diverse range of readers, from neuroscience to engineering, to a common technical understanding. The description of hearing as signal-processing algorithms is supported by corresponding open-source code, for which the book serves as motivating documentation.

Keyword Spotting for Google Assistant Using Contextual Speech Recognition (ASRU 2017; Link)

Abstract:
We present a novel approach for improving overall quality of keyword spotting using contextual automatic speech recognition (ASR) system. On voice-activated devices with limited resources, it is common that a keyword spotting system is run on the device in order to detect a trigger phrase (e.g. “ok google”) and decide which audio should be sent to the server (to be transcribed by the ASR system and processed to generate a response to the user). Due to limited resources on a device, the device keyword spotting system might introduce false accepts (FAs) and false rejects (FRs) that can cause a negative user experience. We describe a system that uses server-side contextual ASR and dynamic classes for improved keyword spotting. We show that this method can significantly reduce FA rates (by 89%) while minimally increasing FR rate (0.15%). Furthermore, we show that this system helps reduce Word Error Rate (WER) (by 10% to 50% relative, on different test sets) and allows users to speak seamlessly, without pausing between the trigger phrase and the command.

Learning to Attend, Copy, and Generate for Session-Based Query Suggestion (CIKM 2017; Link)

Abstract:
Users try to articulate their complex information needs during search sessions by reformulating their queries. In order to make this process more effective, search engines provide related queries to help users to specify the information need in their search process. In this paper, we propose a customized sequence-to-sequence model for session-based query suggestion.In our model, we employ a query-aware attention mechanism to capture the structure of the session context. This enables us to control the scope of the session from which we infer the suggested next query, which helps not only handle the noisy data but also automatically detect session boundaries. Furthermore, we observe that based on user query reformulation behavior, a large portion of terms of a query in a session is retained from the previously submitted queries in the same session and consists of mostly infrequent or unseen terms that are usually not included in the vocabulary. We therefore empower the decoder of our model to access the source words from the session context during decoding by incorporating a copy mechanism. Moreover, we propose evaluation metrics to assess the quality of the generative models for query suggestion. We conduct an extensive set of experiments and analysis. The results suggest that our model outperforms the baselines both in terms of the generating queries and scoring candidate queries for the task of query suggestion.

Learning to Skim Text (ACL 2017; Link)

Abstract:
Recurrent Neural Networks are showing much promise in many sub-areas of natural language processing, ranging from document classification to machine translation to automatic question answering. Despite their promise, many recurrent models have to read the whole text word by word, making it slow to handle long documents. For example, it is difficult to use a recurrent network to read a book and answer questions about it. In this paper, we present an approach of reading text while skipping irrelevant information if needed. The underlying model is a recurrent network that learns how far to jump after reading a few words of the input text. We employ a standard policy gradient method to train the model to make discrete jumping decisions. In our benchmarks on four different tasks, including number prediction, sentiment analysis, news article classification and automatic Q\&A, our proposed model, a modified LSTM with jumping, is up to 6 times faster than the standard sequential LSTM, while maintaining the same or even better accuracy.

Natural Language Processing with Small Feed-Forward Networks (EMNLP. ACL 2017; Link)

Abstract:
We show that small and shallow feedforward neural networks can achieve near state-of-the-art results on a range of unstructured and structured language processing tasks while being considerably cheaper in memory and computational requirements than deep recurrent models. Motivated by resource-constrained environments like mobile phones, we showcase simple techniques for obtaining such small neural network models, and investigate different tradeoffs when deciding how to allocate a small memory budget.

Predicting Latent Structured Intents from Shopping Queries (WWW, ACM 2017; Link)

Abstract:
In online shopping, users usually express their intent through search queries. However, these queries are often ambiguous. For example, it is more likely (and easier) for users to write a query like “high-end bike” than “21 speed carbon frames jamis or giant road bike”. It is challenging to interpret these ambiguous queries and thus search result accuracy suffers. A user oftentimes needs to go through the frustrating process of refining search queries or self-teaching from possibly unstructured information. However, shopping is indeed a structured domain, that is composed of category hierarchy, brands, product lines, features, etc. It would be much better if a shopping site could understand users’ intent through this structure, present organized information, and then find the items with the right categories, brands or features. In this paper we study the problem of inferring the latent intent from

unstructured queries and mapping them to structured attributes. We present a novel framework that jointly learns this knowledge from user consumption behaviors and product metadata. We present a hybrid Long Shortterm Memory (LSTM) [10] joint model that is accurate and robust, even though user queries are noisy and product catalog is rapidly growing. Our study is conducted on a largescale dataset from Google Shopping, that is composed of millions of items and user queries along with their click responses. Extensive qualitative and quantitative evaluation shows that the proposed model is more accurate, concise, and robust than multiple possible alternatives. In terms of information retrieval (IR) performance, our model is able to improve the quality of current Google Shopping production system, which is a very strong baseline.

Google DeepMing Technologies

PaLM 2 - Next Gen LLM (here)

We introduce PaLM 2, a new state-of-the-art language model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM. PaLM 2 is a Transformer-based model trained using a mixture of objectives. Through extensive evaluations on English and multilingual language, and reasoning tasks, we demonstrate that PaLM 2 has significantly improved quality on downstream tasks across different model sizes, while simultaneously exhibiting faster and more efficient inference compared to PaLM. This improved efficiency enables broader deployment while also allowing the model to respond faster, for a more natural pace of interaction. PaLM 2 demonstrates robust reasoning capabilities exemplified by large improvements over PaLM on BIG-Bench and other reasoning tasks. PaLM 2 exhibits stable performance on a suite of responsible AI evaluations, and enables inference-time control over toxicity without additional overhead or impact on other capabilities. Overall, PaLM 2 achieves state-of-the-art performance across a diverse set of tasks and capabilities.

PaLM 2 is our next generation large language model that builds on Google's legacy of breakthrough research in machine learning and responsible AI.

It excels at advanced reasoning tasks, including code and math, classification and question answering, translation and multilingual proficiency, and natural language generation better than our previous state-of-the-art LLMs, including PaLM. It can accomplish these tasks because of the way it was built – bringing together compute-optimal scaling, an improved dataset mixture, and model architecture improvements.

PaLM 2 is grounded in Google's approach to building and deploying AI responsibly. All versions of PaLM 2 are evaluated rigorously for potential harms and biases, capabilities and downstream uses in research and in-product applications.

SynthID - Identify AI Generated Images (here)

We're beta launching SynthID, a tool for watermarking and identifying AI-generated images. SynthID is being released to a limited number of Vertex AI customers using Imagen, one of our latest text-to-image models that uses input text to create photorealistic images. With this tool, users can embed an imperceptible digital watermark into their AI-generated images and identify if Imagen was used for generating the image, or even part of the image.

SynthID uses two deep learning models — one for watermarking and another for identifying — which were trained together on a diverse set of images. The combined model is optimised on a range of objectives, including correctly identifying watermarked content and improving imperceptibility by visually aligning the watermark to the original content.

Imagen - Unprecedented Photorealism × Deep Level of Language Understanding (here)

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment.

Phenaki - Realistic Video Generation from Open-Domain Textual Descriptions (here)

We present Phenaki, a model that can synthesize realistic videos from textual prompt sequences. Generating videos from text is particularly challenging due to various factors, such as high computational cost, variable video lengths, and limited availability of high quality text-video data.

To address the first two issues, Phenaki leverages its two main components:

- An **encoder-decoder model** that compresses videos to discrete embeddings, or tokens, with a tokenizer that can work with variable-length videos thanks to its use of causal attention in time.

- A **transformer model** that translates text embeddings to video tokens: we use a bi-directional masked transformer conditioned on pre-computed text tokens to generate video tokens from text, which are subsequently de-tokenized to create the actual video.

AlphaZero and MuZero - Powerful, General AI Systems, that mastered a range of board games and video games (here)

AlphaZero: A dynamic and creative player
AlphaZero represents a crucial step towards creating more general systems. It taught itself, from scratch, to master the board games of chess, shogi, and Go. In doing so, it became the strongest player in history for each. The system is the successor to AlphaGo, the first AI to defeat a professional human Go player and one that inspired a new era of AI advances.

Unlike AlphaGo, which learned to play Go by analyzing millions of moves from amateur games, AlphaZero's neural network was only given the rules of each game. It then learned each game by playing itself millions of times. Through a process of trial and error, called reinforcement learning, the system learned to select the most promising moves and boost its chances of winning.

AlphaZero mastered chess in just 9 hours. Shogi in 12 hours. And Go in 13 days. In each game, it learned to play with a unique and creative style.

MuZero: AI that can plan
MuZero goes one step further than AlphaZero. Without being told the rules of any game, MuZero matches AlphaZero's level of performance in Go, chess and shogi, and also learns to master a suite of visually complex Atari games.

It does this by learning a model of its environment, such as the game it's playing. MuZero then uses that model to plan the best course of action. Crucially, it only models three aspects of its environment that are important to its decision-making process - how good is the current position? Which action is the best to take? And how good was the last action?

These are all learned using a deep neural network and are all that is needed for MuZero to understand what happens when it takes an action and to plan accordingly.

AlphaGo - Mastered the ancient game of Go (here)

Go was long considered a grand challenge for AI. The game is a googol times more complex than chess — with an astonishing 10 to the power of 170 possible board configurations. That's more than the number of atoms in the known universe.

The strongest Go computer programs only achieved the level of human amateurs, despite decades of work. Standard AI methods struggled to assess the sheer number of possible moves and lacked the creativity and intuition of human players.

We created AlphaGo, an AI system that combines deep neural networks with advanced search algorithms. One neural network — known as the "policy network" — selects the next move to play. The other neural network — the "value network" — predicts the winner of the game.
Initially, we introduced AlphaGo to numerous amateur games of Go so the system could learn how humans play the game. Then we instructed AlphaGo to play against different versions of itself thousands of times, each time learning from its mistakes — a method known as reinforcement learning. Over time, AlphaGo improved and became a better player.

PaLM-SayCan - Cutting-edge robotics algorithm (here)

A robotics algorithm that combines the understanding of language models with the real-world capabilities of a helper robot.

Universal Speech Model (USM) - Towards Automatic Speech Recognition for All (here)

Universal Speech Model (USM) is a family of state-of-the-art speech models with 2B parameters trained on 12 million hours of speech and 28 billion sentences of text, spanning 300+ languages. USM, which is for use in YouTube (e.g., for closed captions), can perform automatic speech recognition (ASR) on widely-spoken languages like English and Mandarin, but also languages like Punjabi, Assamese, Santhali, Balinese, Shona, Malagasy, Luganda, Luo, Bambara, Soga, Maninka, Xhosa, Akan, Lingala, Chichewa, Nkore, Nzema to name a few. Some of these languages are spoken by fewer than twenty million people, making it very hard to find the necessary training data.

We demonstrate that utilizing a large unlabeled multilingual dataset to pre-train the encoder of our model and fine-tuning on a smaller set of labeled data enables us to recognize these under-represented languages. Moreover, our model training process is effective for adapting to new languages and data.

WaveNet - Generate Natural-Sounding Speech (here)

Introduced in 2016, WaveNet was one of the first AI models to generate natural-sounding speech. Since then, it has inspired research, products, and applications in Google — and beyond.

WaveNet is a generative model trained on human speech samples. It creates waveforms of speech patterns by predicting which sounds are most likely to follow each other, each built one sample at a time, with up to 24,000 samples per second of sound.

The model incorporates natural-sounding elements, such as lip-smacking and breathing patterns. And includes vital layers of communication like intonation, accents, emotion — delivering a richness and depth to computer-generated voices.
Using a technique called distillation — transferring knowledge from a larger to smaller model — we reengineered WaveNet to run 1,000 times faster than our research prototype, creating one second of speech in just 50 milliseconds.

In parallel, we also developed WaveRNN — a simpler, faster, and more computationally efficient model that could run on devices, like mobile phones, rather than in a data center.

AlphaCode - Competitive Programming (here)

The problem-solving abilities required to excel at these competitions are beyond the capabilities of existing AI systems. However, by combining advances in large-scale transformer models (that have recently shown promising abilities to generate code) with large-scale sampling and filtering, we've made significant progress in the number of problems we can solve. We pre-train our model on selected public GitHub code and fine-tune it on our relatively small competitive programming dataset.

At evaluation time, we create a massive amount of C++ and Python programs for each problem, orders of magnitude larger than previous work. Then we filter, cluster, and rerank those solutions to a small set of 10 candidate programs that we submit for external assessment. This automated system replaces competitors' trial-and-error process of debugging, compiling, passing tests, and eventually submitting.

With the permission of Codeforces, we evaluated AlphaCode by simulating participation in 10 recent contests. The impressive work of the competitive programming community has created a domain where it's not possible to solve problems through shortcuts like duplicating solutions seen before or trying out every potentially related algorithm. Instead, our model must create novel and interesting solutions.

Overall, AlphaCode placed at approximately the level of the median competitor. Although far from winning competitions, this result represents a substantial leap in AI problem-solving

capabilities and we hope that our results will inspire the competitive programming community.

AlphaTensor - First AI System for discovering novel, efficient, and provably correct algorithms (here)

We introduce AlphaTensor, the first artificial intelligence (AI) system for discovering novel, efficient, and provably correct algorithms for fundamental tasks such as matrix multiplication. This sheds light on a 50-year-old open question in mathematics about finding the fastest way to multiply two matrices.

This paper is a stepping stone in DeepMind's mission to advance science and unlock the most fundamental problems using AI. Our system, AlphaTensor, builds upon AlphaZero, an agent that has shown superhuman performance on board games, like chess, Go and shogi, and this work shows the journey of AlphaZero from playing games to tackling unsolved mathematical problems for the first time.

AlphaStar - AI to Master Real-Time Strategy Game StarCraft II using multi-agent reinforcement learning (here)

AlphaStar is the first AI to reach the top league of a widely popular esport without any game restrictions. This January, a preliminary version of AlphaStar challenged two of the world's top players in StarCraft II, one of the most enduring and popular real-time strategy video games of all time. Since then, we have taken on a much greater challenge: playing the full game at a Grandmaster level under professionally approved conditions.

We chose to use general-purpose machine learning techniques – including neural networks, self-play via reinforcement learning, multi-agent learning, and imitation learning – to learn directly from game data with general purpose techniques. Using the advances described in our Nature paper, AlphaStar was ranked above 99.8% of active players on Battle.net, and achieved a Grandmaster level for all three StarCraft II races: Protoss, Terran, and Zerg. We expect these methods could be applied to many other domains.

Learning-based systems and self-play are elegant research concepts which have facilitated remarkable advances in artificial intelligence. In 1992, researchers at IBM developed TD-Gammon, combining a learning-based system with a neural network to play the game of backgammon. Instead of playing according to hard-coded rules or heuristics, TD-Gammon was designed to use reinforcement learning to figure out, through trial-and-error, how to play the game in a way that maximises its probability of winning. Its developers used the notion of self-play to make the system more robust: by playing against versions of itself, the system grew increasingly proficient at the game. When combined, the notions of learning-based systems and self-play provide a powerful paradigm of open-ended learning.

Many advances since then have demonstrated that these approaches can be scaled to progressively challenging domains. For example, AlphaGo and AlphaZero established that it was possible for a system to learn to achieve superhuman performance at Go, chess, and shogi, and OpenAI Five and DeepMind's FTW demonstrated the power of self-play in the modern games of Dota 2 and Quake III.

Magenta: Making Music and art with Machine Learning (here)

An open source research project exploring the role of machine learning as a tool in the creative process.

Project Euphonia - Helping people be better understood (here)

Project Euphonia is a Google Research initiative focused on helping people with non-standard speech be better understood. The approach is centered on analyzing speech recordings to better train speech recognition models.

For millions of people around the world whose speech is difficult for others to understand, face-to-face communication can be very challenging. Using voice-activated technologies can be frustrating, too. While tools like Google Home or Google Assistant can help people call someone, adjust lighting, or play a favorite song, they may not work as well for those with non-standard speech.