Meta AI

Research Areas: Conversational AI, Natural Language Processing, Reinforcement Learning, Responsible AI
Years: 2023, 2022, 2021, 2020, 2019, 2018, 2017
Number of Research Papers Evaluated: 358
Number of Research Papers After Evaluation: 142
Author: Dhruv Awasthi

Q-Pensieve: Boosting Sample Efficiency of Multi-Objective RL Through Memory Sharing of Q-Snapshots (ICLR, Oct 2023; Link)

Abstract:
Many real-world continuous control problems are in the dilemma of weighing the pros and cons, multi-objective reinforcement learning (MORL) serves as a generic framework of learning control policies for different preferences over objectives. However, the existing MORL methods either rely on multiple passes of explicit search for finding the Pareto front and therefore are not sample-efficient, or utilizes a shared policy network for coarse knowledge sharing among policies. To boost the sample efficiency of MORL, we propose -Pensieve, a policy improvement scheme that stores a collection of -snapshots to jointly determine the policy update direction and thereby enables data sharing at the policy level. We show that -Pensieve can be naturally integrated with soft policy iteration with convergence guarantee. To substantiate this concept, we propose the technique of replay buffer, which stores the learned -networks from the past iterations, and arrive at a practical actor-critic implementation. Through extensive experiments and an ablation study, we demonstrate that with much fewer samples, the proposed algorithm can outperform the benchmark MORL methods on a variety of MORL benchmark tasks.

Effective Long-Context Scaling of Foundation Models (Meta, Sep 2023; Link)

Abstract:
We present a series of long-context LLMs that support effective context windows of up to 32,768 tokens. Our model series are built through continual pretraining from LLAMA 2 with longer training sequences and on a dataset where long texts are upsampled. We perform extensive evaluation on language modeling, synthetic context probing tasks, and a wide range of research benchmarks. On research benchmarks, our models achieve consistent improvements on most regular tasks and significant improvements on long-context tasks over LLAMA 2. Notably, with a cost-effective instruction tuning procedure that does not require human-annotated long instruction data, the 70B variant can already surpass gpt-3.5-turbo-16k's overall performance on a suite of long-context tasks. Alongside these

results, we provide an in-depth analysis on the individual components of our method. We delve into LLAMA's position encodings and discuss its limitation in modeling long dependencies. We also examine the impact of various design choices in the pretraining process, including the data mix and the training curriculum of sequence lengths – our ablation experiments suggest that having abundant long texts in the pretrain dataset is not the key to achieving strong performance, and we empirically verify that long context continual pretraining is more efficient and similarly effective compared to pretraining from scratch with long sequences.

Code Llama: Open Foundation Models for Code (Meta, Aug 2023; Link)

Abstract:
We release Code Llama, a family of large language models for code based on Llama 2 providing state-of-the-art performance among open models, infilling capabilities, support for large input contexts, and zero-shot instruction following ability for programming tasks. We provide multiple flavors to cover a wide range of applications: foundation models (Code Llama), Python specializations (Code Llama - Python), and instruction-following models (Code Llama - Instruct) with 7B, 13B and 34B parameters each. All models are trained on sequences of 16k tokens and show improvements on inputs with up to 100k tokens. 7B and 13B Code Llama and Code Llama - Instruct variants support infilling based on surrounding content. Code Llama reaches state-of-the-art performance among open models on several code benchmarks, with scores of up to 53% and 55% on HumanEval and MBPP, respectively. Notably, Code Llama - Python 7B outperforms Llama 2 70B on HumanEval and MBPP, and all our models outperform every other publicly available model on MultiPL-E. We release Code Llama under a permissive license that allows for both research and commercial use.

SeamlessM4T—Massively Multilingual & Multimodal Machine Translation (Meta, Aug 2023; Link)

Abstract:
What does it take to create the Babel Fish, a tool that can help individuals translate speech between any two languages? While recent breakthroughs in text-based models have pushed machine translation coverage beyond 200 languages, unified speech-to-speech translation models have yet to achieve similar strides. More specifically, conventional speech-to-speech translation systems rely on cascaded systems composed of multiple subsystems performing translation progressively, putting scalable and high-performing unified speech translation systems out of reach. To address these gaps, we introduce SeamlessM4T—Massively Multilingual & Multimodal Machine Translation—a single model that supports speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages. To build

this, we used 1 million hours of open speech audio data to learn self-supervised speech representations with w2v-BERT 2.0. Subsequently, we created a multimodal corpus of automatically aligned speech translations, dubbed SeamlessAlign. Filtered and combined with human labeled and pseudo-labeled data (totaling 406,000 hours), we developed the first multilingual system capable of translating from and into English for both speech and text. On Fleurs, SeamlessM4T sets a new standard for translations into multiple target languages, achieving an improvement of 20% BLEU over the previous state-of-the-art in direct speech-to-text translation. Compared to strong cascaded models, SeamlessM4T improves the quality of into-English translation by 1.3 BLEU points in speech-to-text and by 2.6 ASR-BLEU points in speech-to-speech. On CVSS and compared to a 2-stage cascaded model for speech-to-speech translation, SeamlessM4T-Large's performance is stronger by 58%. Preliminary human evaluations of speech-to-text translation outputs evinced similarly impressive results; for translations from English, XSTS scores for 24 evaluated languages are consistently above 4 (out of 5). For into English directions, we see significant improvement over WhisperLarge-v2's baseline for 7 out of 24 languages. To further evaluate our system, we developed Blaser 2.0, which enables evaluation across speech and text with similar accuracy compared to its predecessor when it comes to quality estimation. Tested for robustness, our system performs better against background noises and speaker variations in speech-to-text tasks (average improvements of 38% and 49%, respectively) compared to the current state-of-the-art model. Critically, we evaluated SeamlessM4T on gender bias and added toxicity to assess translation safety. Compared to the state-of-the-art, we report up to 63% of reduction in added toxicity in our translation outputs. Finally, all contributions in this work—including models, inference code, finetuning recipes backed by our improved modeling toolkit Fairseq2, and metadata to recreate the unfiltered 470,000 hours of SeamlessAlign — are open-sourced and accessible at this link.

SONAR: Sentence-Level Multimodal and Language-Agnostic Representations (Meta, Aug 2023; Link)

Abstract:
We introduce SONAR, a new multilingual and multimodal fixed-size sentence embedding space. Our single text encoder, covering 200 languages, substantially outperforms existing sentence embeddings such as LASER3 and LabSE on the xsim and xsim++ multilingual similarity search tasks. Speech segments can be embedded in the same SONAR embedding space using language-specific speech encoders trained in a teacher-student setting on speech transcription data. Our encoders outperform existing speech encoders on similarity search tasks. We also provide a text decoder for 200 languages, which allows us to perform text-to-text and speech-to-text machine translation, including for zero-shot language and modality combinations. Our text-to-text results are competitive compared to the state-of-the-art NLLB 1B model, despite the fixed-size bottleneck representation. Our zero-shot speech-to-text translation results compare favorably with strong supervised baselines such as Whisper.

MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation

(INTERSPEECH, Aug 2023; Link)

Abstract:
We introduce MuAViC, a multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation providing 1200 hours of audio-visual speech in 9 languages. It is fully transcribed and covers 6 English-to-X translation as well as 6 X-to-English translation directions. To the best of our knowledge, this is the first open benchmark for audio-visual speech-to-text translation and the largest open benchmark for multilingual audio-visual speech recognition. Our baseline results show that MuAViC is effective for building noise-robust speech recognition and translation models. We make the corpus available at https://github.com/facebookresearch/muavic.

Multi-Head State Space Model for Speech Recognition

(INTERSPEECH, Aug 2023; Link)

Abstract:
State space models (SSMs) have recently shown promising results on small-scale sequence and language modelling tasks, rivalling and outperforming many attention-based approaches. In this paper, we propose a multi-head state space (MH-SSM) architecture equipped with special gating mechanisms, where parallel heads are taught to learn local and global temporal dynamics on sequence data. As a drop-in replacement for multi-head attention in transformer encoders, this new model significantly outperforms the transformer transducer on the LibriSpeech speech recognition corpus. Furthermore, we augment the transformer block with MH-SSMs layers, referred to as the Stateformer, achieving state-of-the-art performance on the LibriSpeech task, with word error rates of 1.76%/4.37% on the development and 1.91%/4.36% on the test sets without using an external language model.

Llama 2: Open Foundation and Fine-Tuned Chat Models (arxiv, July 2023; Link)

Abstract:
In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama 2-Chat, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closedsource models. We provide a detailed description of our approach to fine-tuning and safety

improvements of Llama 2-Chat in order to enable the community to build on our work and contribute to the responsible development of LLMs.

Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning (Meta, Jul 2023; Link)

Abstract:
We present CM3Leon (pronounced “Chameleon”), a retrieval-augmented, tokenbased, decoder-only multi-modal language model capable of generating and infilling both text and images. CM3Leon uses the CM3 multi-modal architecture but additionally shows the extreme benefits of scaling up and tuning on more diverse instruction-style data. It is the first multi-modal model trained with a recipe adapted from text-only language models, including a large-scale retrieval-augmented pretraining stage and a second multi-task supervised fine-tuning (SFT) stage. It is also a general purpose model that can do both text-to-image and image-to text generation, allowing us to introduce self-contained contrastive decoding methods that produce high-quality outputs. Extensive experiments demonstrate that this recipe is highly effective for multi-modal models. CM3Leon achieves state-of-theart performance in text-to-image generation with 5x less training compute than comparable methods (zero-shot MS-COCO FID of 4.88). After SFT, CM3Leon can also demonstrate unprecedented levels of controllability in tasks ranging from language-guided image editing to image-controlled generation and segmentation.

StructDiffusion: Language-Guided Creation of Physically-Valid Structures using Unseen Objects (Robotics Science and Systems, Jul 2023; Link)

Abstract:
Robots operating in human environments must be able to rearrange objects into semantically-meaningful configurations, even if these objects are previously unseen. In this work, we focus on the problem of building physically-valid structures without step-by-step instructions. We propose StructDiffusion, which combines a diffusion model and an object-centric transformer to construct structures given partial-view point clouds and high-level language goals, such as "set the table". Our method can perform multiple challenging language-conditioned multi-step 3D planning tasks using one model. StructDiffusion even improves the success rate of assembling physically-valid structures out of unseen objects by on average 16% over an existing multi-modal transformer model trained on specific structures. We show experiments on held-out objects in both simulation and on real-world rearrangement tasks. Importantly, we show how integrating both a diffusion model and a collision-discriminator model allows for improved generalization over other methods when rearranging previously-unseen objects. For videos and additional results, see our website: https://structdiffusion.github.io/.

Galactic: Scaling End-to-End Reinforcement Learning for Rearrangement at 100k Steps-Per-Second (CVPR, Jun 2023; Link)

Abstract:
We present Galactic, a large-scale simulation and reinforcement-learning (RL) framework for robotic mobile manipulation in indoor environments. Specifically, a Fetch robot (equipped with a mobile base, 7DoF arm, RGBD camera, egomotion, and onboard sensing) is spawned in a home environment and asked to rearrange objects – by navigating to an object, picking it up, navigating to a target location, and then placing the object at the target location. Galactic is fast. In terms of simulation speed (rendering + physics), Galactic achieves over 421,000 steps-per-second (SPS) on an 8-GPU node, which is 54x faster than Habitat 2.0 [55] (7699 SPS). More importantly, Galactic was designed to optimize the entire rendering+physics+RL interplay since any bottleneck in the interplay slows down training. In terms of simulation+RL speed (rendering + physics + inference + learning), Galactic achieves over 108,000 SPS, which 88x faster than Habitat 2.0 (1243 SPS). These massive speed-ups not only drastically cut the wall-clock training time of existing experiments, but also unlock an unprecedented scale of new experiments. First, Galactic can train a mobile pick skill to > 80% accuracy in under 16 minutes, a 100x speedup compared to the over 24 hours it takes to train the same skill in Habitat 2.0. Second, we use Galactic to perform the largest-scale experiment to date for rearrangement using 5B steps of experience in 46 hours, which is equivalent to 20 years of robot experience. This scaling results in a single neural network composed of task-agnostic components achieving 85% success in GeometricGoal rearrangement, compared to 0% success reported in Habitat 2.0 for the same approach. The code is available at github.com/facebookresearch/galactic .

Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale (Meta, Jun 2023; Link)

Abstract:
Large-scale generative models such as GPT and DALL-E have revolutionized natural language processing and computer vision research. These models not only generate high fidelity text or image outputs, but are also generalists which can solve tasks not explicitly taught. In contrast, speech generative models are still primitive in terms of scale and task generalization. In this paper, we present Voicebox, the most versatile text-guided generative model for speech at scale. Voicebox is a non-autoregressive flow-matching model trained to infill speech, given audio context and text, trained on over 50K hours of speech that are neither filtered nor enhanced. Similar to GPT, Voicebox can perform many different tasks through in-context learning, but is more flexible as it can also condition on future context. Voicebox can be used for mono or cross-lingual zero-shot text-to-speech synthesis, noise removal, content editing, style conversion, and diverse sample generation. In particular, Voicebox outperforms the state-of-the-art zero-shot TTS model VALL-E on both intelligibility (5.9% vs 1.9% word error rates) and audio similarity (0.580 vs 0.681) while being up to 20 times faster. See voicebox.metademolab.com for a demo of the model

PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English (ACL, Jun 2023; Link)

Abstract:
Privacy policies provide individuals with information about their rights and how their personal information is handled. Natural language understanding (NLU) technologies can support individuals and practitioners to understand better privacy practices described in lengthy and complex documents. However, existing efforts that use NLU technologies are limited by processing the language in a way exclusive to a single task focusing on certain privacy practices. To this end, we introduce the Privacy Policy Language Understanding Evaluation (PLUE) benchmark, a multi-task benchmark for evaluating the privacy policy language understanding across various tasks. We also collect a large corpus of privacy policies to enable privacy policy domain-specific language model pre-training. We evaluate several generic pre-trained language models and continue pre-training them on the collected corpus. We demonstrate that domain-specific continual pre-training offers performance improvements across all tasks.

Handling the Alignment forWake Word Detection: A Comparison Between Alignment-Based, Alignment-Free and Hybrid Approaches (INTERSPEECH, Jun 2023; Link)

Abstract:
Wake word detection exists in most intelligent homes and portable devices. It offers these devices the ability to “wake up” when summoned at a low cost of power and computing. This paper focuses on understanding alignment’s role in developing a wake-word system that answers a generic phrase. We discuss three approaches. The first is alignment-based, where the model is trained with frame-wise cross-entropy. The second is alignment-free, where the model is trained with CTC. The third, proposed by us, is a hybrid solution in which the model is trained with a small set of aligned data and then tuned with a sizeable unaligned dataset. We compare the three approaches and evaluate the impact of the different aligned-to-unaligned ratios for hybrid training. Our results show that the alignment-free system performs better than the alignment-based for the target operating point, and with a small fraction of the data (20%), we can train a model that complies with our initial constraints.

Scaling Speech Technology to 1,000+ Languages (NeurIPS, May 2023; Link)

Abstract:
Expanding the language coverage of speech technology has the potential to improve access to information for many more people. However, current speech technology is restricted to

about one hundred languages which is a small fraction of the over 7,000 languages spoken around the world. The Massively Multilingual Speech (MMS) project increases the number of supported languages by 10-40x, depending on the task. The main ingredients are a new dataset based on readings of publicly available religious texts and effectively leveraging self-supervised learning. We built pre-trained wav2vec 2.0 models covering 1,406 languages, a single multilingual automatic speech recognition model for 1,107 languages, speech synthesis models for the same number of languages, as well as a language identification model for 4,017 languages. Experiments show that our multilingual speech recognition model more than halves the word error rate of Whisper on 54 languages of the FLEURS benchmark while being trained on a small fraction of the labeled data.

MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations (ICLR, May 2023; Link)

Abstract:
Poor sample efficiency continues to be the primary challenge for deployment of deep Reinforcement Learning (RL) algorithms for real-world applications, and in particular for visuo-motor control. Model-based RL has the potential to be highly sample efficient by concurrently learning a world model and using synthetic rollouts for planning and policy improvement. However, in practice, sample-efficient learning with model-based RL is bottlenecked by the exploration challenge. In this work, we find that leveraging just a handful of demonstrations can dramatically improve the sample-efficiency of model-based RL. Simply appending demonstrations to the interaction dataset, however, does not suffice. We identify key ingredients for leveraging demonstrations in model learning -- policy pretraining, targeted exploration, and oversampling of demonstration data -- which forms the three phases of our model-based RL framework. We empirically study three complex visuo-motor control domains and find that our method is 150%-250% more successful in completing sparse reward tasks compared to prior approaches in the low data regime (100K interaction steps, 5 demonstrations).

PIRLNav: Pretraining with Imitation and RL Finetuning for ObjectNav (CVPR, Mar 2023; Link)

Abstract:
We study ObjectGoal Navigation -- where a virtual robot situated in a new environment is asked to navigate to an object. Prior work has shown that imitation learning (IL) using behavior cloning (BC) on a dataset of human demonstrations achieves promising results. However, this has limitations -- 1) BC policies generalize poorly to new states, since the training mimics actions not their consequences, and 2) collecting demonstrations is expensive. On the other hand, reinforcement learning (RL) is trivially scalable, but requires careful reward engineering to achieve desirable behavior. We present PIRLNav, a two-stage learning scheme for BC pretraining on human demonstrations followed by RL-finetuning.

This leads to a policy that achieves a success rate of 65.0% on ObjectNav (+5.0% absolute over previous state-of-the-art). Using this BC→RL training recipe, we present a rigorous empirical analysis of design choices. First, we investigate whether human demonstrations can be replaced with `free' (automatically generated) sources of demonstrations, e.g. shortest paths (SP) or task-agnostic frontier exploration (FE) trajectories. We find that BC→RL on human demonstrations outperforms BC→RL on SP and FE trajectories, even when controlled for same BC-pretraining success on train, and even on a subset of val episodes where BC-pretraining success favors the SP or FE policies. Next, we study how RL-finetuning performance scales with the size of the BC pretraining dataset. We find that as we increase the size of BC-pretraining dataset and get to high BC accuracies, improvements from RL-finetuning are smaller, and that 90% of the performance of our best BC→RL policy can be achieved with less than half the number of BC demonstrations. Finally, we analyze failure modes of our ObjectNav policies, and present guidelines for further improving them.

LLaMA: Open and Efficient Foundation Language Models (ArXiV, Feb 2023; Link)

Abstract:
We introduce LLaMA, a collection of founda- tion language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla- 70B and PaLM-540B. We release all our models to the research community.

UNIREX: A Unified Learning Framework For Language Model Rationale Extraction (NACCL, Feb 2023; Link)

Abstract:
An extractive rationale explains a language model's (LM's) prediction on a given task instance by highlighting the text inputs that most influenced the prediction. Ideally, rationale extraction should be faithful (reflective of LM's actual behavior) and plausible (convincing to humans), without compromising the LM's (i.e., task model's) task performance. Although attribution algorithms and select-predict pipelines are commonly used in rationale extraction, they both rely on certain heuristics that hinder them from satisfying all three desiderata. In light of this, we propose UNIREX, a flexible learning framework which generalizes rationale extractor optimization as follows: (1) specify architecture for a learned rationale extractor; (2) select explainability objectives (i.e., faithfulness and plausibility criteria); and (3) jointly train the task model and rationale extractor on the task using selected objectives. UNIREX enables replacing prior works' heuristic design choices with a generic learned rationale extractor in (1) and optimizing it for all three desiderata in (2)-(3). To facilitate comparison between methods w.r.t. multiple desiderata, we introduce the Normalized Relative Gain

(NRG) metric. Across five English text classification datasets, our best UNIREX configuration outperforms baselines by an average of 32.9% NRG. Plus, we find that UNIREX-trained rationale extractors' faithfulness can even generalize to unseen datasets and tasks.

Staircase Attention for Recurrent Processing of Sequences (NeurIPS, Dec 2022; Link)

Abstract:
Attention mechanisms have become a standard tool for sequence modeling tasks, in particular by stacking self-attention layers over the entire input sequence as in the Transformer architecture. In this work we introduce a novel attention procedure called staircase attention that, unlike self-attention, operates across the sequence (in time) recurrently processing the input by adding another step of processing. A step in the staircase comprises of backward tokens (encoding the sequence so far seen) and forward tokens (ingesting a new part of the sequence). Thus our model can trade off performance and compute, by increasing the amount of recurrence through time and depth. Staircase attention is shown to be able to solve tasks that involve tracking that conventional Transformers cannot, due to this recurrence. Further, it is shown to provide improved modeling power for the same size model (number of parameters) compared to self-attentive Transformers on large language modeling and dialogue tasks, yielding significant perplexity gains.

Efficient Self-supervised Learning with Contextualized Target Representations for Vision, Speech and Language (ArXiV, Dec 2022; Link)

Abstract:
Current self-supervised learning algorithms are often modality-specific and require large amounts of computational resources. To address these issues, we increase the training efficiency of data2vec, a learning objective that generalizes across several modalities. We do not encode masked tokens, use a fast convolutional decoder and amortize the effort to build teacher representations. data2vec 2.0 benefits from the rich contextualized target representations introduced in data2vec which enable a fast self-supervised learner. Experiments on ImageNet-1K image classification show that data2vec 2.0 matches the accuracy of Masked Autoencoders in 16.4x lower pre-training time, on Librispeech speech recognition it performs as well as wav2vec 2.0 in 10.6x less time, and on GLUE natural language understanding it matches a retrained RoBERTa model in half the time. Trading some speed for accuracy results in ImageNet-1K top-1 accuracy of 86.8\% with a ViT-L model trained for 150 epochs.

"That's so cute!": The CARE Dataset for Affective Response Detection (CoNLL, Dec 2020; Link)

Abstract:
Social media plays an increasing role in our communication with friends and family, and in our consumption of entertainment and information. Hence, to design effective ranking functions for posts on social media, it would be useful to predict the affective responses of a post (e.g., whether it is likely to elicit feelings of entertainment, inspiration, or anger). Similar to work on emotion detection (which focuses on the affect of the publisher of the post), the traditional approach to recognizing affective response would involve an expensive investment in human annotation of training data. We create and publicly release CARE DB, a dataset of 230k social media post annotations according to seven affective responses using the Common Affective Response Expression (CARE) method. The CARE method is a means of leveraging the signal that is present in comments that are posted in response to a post, providing high-precision evidence about the affective response to the post without human annotation. Unlike human annotation, the annotation process we describe here can be iterated upon to expand the coverage of the method, particularly for new affective responses. We present experiments that demonstrate that the CARE annotations compare favorably with crowdsourced annotations. Finally, we use CARE DB to train competitive BERT-based models for predicting affective response as well as emotion detection, demonstrating the utility of the dataset for related tasks.

The Curious Case of Absolute Position Embeddings (EMNLP, Dec 2020; Link)

Abstract:
Transformer language models encode the notion of word order using positional information. Most commonly, this positional information is represented by absolute position embeddings (APEs), that are learned from the pretraining data. However, in natural language, it is not absolute position that matters, but relative position, and the extent to which APEs can capture this type of information has not been investigated. In this work, we observe that models trained with APE over-rely on positional in- formation to the point that they break-down when subjected to sentences with shifted posi- tion information. Specifically, when models are subjected to sentences starting from a non-zero position (excluding the effect of priming), they exhibit noticeably degraded performance on zero- to full-shot tasks, across a range of model families and model sizes. Our findings raise questions about the efficacy of APEs to model the relativity of position information, and invite further introspection on the sentence and word order processing strategies employed by these models.

ToKen: Task Decomposition and Knowledge Infusion for Few-Shot Hate Speech Detection (EMNLP, Nov 2020; Link)

Abstract:
Hate speech detection is complex; it relies on commonsense reasoning, knowledge of stereotypes, and an understanding of social nuance that differs from one culture to the next. It is also difficult to collect a large-scale hate speech annotated dataset. In this work, we frame this problem as a few-shot learning task, and show significant gains with decomposing the task into its "constituent" parts. In addition, we see that infusing knowledge from reasoning datasets (e.g. ATOMIC) improves the performance even further. Moreover, we observe that the trained models generalize to out-of-distribution datasets, showing the superiority of task decomposition and knowledge infusion compared to previously used methods. Concretely, our method outperforms the baseline by 17.83% absolute gain in the 16-shot case.

Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models (NeurIPS, Nov 2022; Link)

Abstract:
Despite their wide adoption, the underlying training and memorization dynamics of very large language models is not well understood. We empirically study exact memorization in causal and masked language modeling, across model sizes and throughout the training process. We measure the effects of dataset size, learning rate, and model size on memorization, finding that larger language models memorize training data faster across all settings. Surprisingly, we show that larger models can memorize a larger portion of the data before over-fitting and tend to forget less throughout the training process. We also analyze the memorization dynamics of different parts of speech and find that models memorize nouns and numbers first; we hypothesize and provide empirical evidence that nouns and numbers act as a unique identifier for memorizing individual training examples. Together, these findings present another piece of the broader puzzle of trying to understand what actually improves as models get bigger.

Autoregressive Search Engines: Generating Substrings as Document Identifiers (ARR/NeurIPS, Oct 2022; Link)

Abstract:
Knowledge-intensive language tasks require NLP systems to both provide the correct answer and retrieve supporting evidence for it in a given corpus. Autoregressive language models are emerging as the de-facto standard for generating answers, with newer and more powerful systems emerging at an astonishing pace. In this paper we argue that all this (and future) progress can be directly applied to the retrieval problem with minimal intervention to

the models' architecture. Previous work has explored ways to partition the search space into hierarchical structures and retrieve documents by autoregressively generating their unique identifier. In this work we propose an alternative that doesn't force any structure in the search space: using all ngrams in a passage as its possible identifiers. This setup allows us to use an autoregressive model to generate and score distinctive ngrams, that are then mapped to full passages through an efficient data structure. Empirically, we show this not only outperforms prior autoregressive approaches but also leads to an average improvement of at least 10 points over more established retrieval solutions for passage-level retrieval on the KILT benchmark, establishing new state-of-the-art downstream performance on some datasets, while using a considerably lighter memory footprint than competing systems. Code and pre-trained models are available at https://github.com/facebookresearch/SEAL.

CCQA: A New Web-Scale Question Answering Dataset for Model Pre-Training (NAACL, Jul 2022; Link)

Abstract:
With the rise of large-scale pre-trained language models, open-domain question-answering (ODQA) has become an important research topic in NLP. Based on the popular pre-training fine-tuning approach, we posit that an additional in-domain pre-training stage using a large-scale, natural, and diverse question-answering (QA) dataset can be beneficial for ODQA. Consequently, we propose a novel QA dataset based on the Common Crawl project in this paper. Using the readily available schema.org annotation, we extract around 130 million multilingual question-answer pairs, including about 60 million English data-points. With this previously unseen number of natural QA pairs, we pre-train popular language models to show the potential of large-scale in-domain pre-training for the task of question-answering. In our experiments, we find that pre-training question-answering models on our Common Crawl Question Answering dataset (CCQA) achieves promising results in zero-shot, low resource and fine-tuned settings across multiple tasks, models and benchmarks.

Learning Accurate Long-term Dynamics for Model-based Reinforcement Learning (CDC, Dec 2021; Link)

Abstract:
Accurately predicting the dynamics of robotic systems is crucial for model-based control and reinforcement learning. The most common way to estimate dynamics is by fitting a one-step ahead prediction model and using it to recursively propagate the predicted state distribution over long horizons. Unfortunately, this approach is known to compound even small prediction errors, making long-term predictions inaccurate. In this paper, we propose a new parametrization to supervised learning on state-action data to stably predict at longer horizons – that we call a trajectory-based model. This trajectory-based model takes an initial state, a future time index, and control parameters as inputs, and directly predicts the state at

the future time index. Experimental results in simulated and real-world robotic tasks show that trajectory-based models yield significantly more accurate long term predictions, improved sample efficiency, and the ability to predict task reward. With these improved prediction properties, we conclude with a demonstration of methods for using the trajectory-based model for control.

Pay Better Attention to Attention: Head Selection in Multilingual and Multi-Domain Sequence Modeling (NeurIPS, Dec 2021; Link)

Abstract:
Multi-head attention has each of the attention heads collect salient information from different parts of an input sequence, making it a powerful mechanism for sequence modeling. Multilingual and multi-domain learning are common scenarios for sequence modeling, where the key challenge is to maximize positive transfer and mitigate negative interference across languages and domains. In this paper, we find that non-selective attention sharing is sub-optimal for achieving good generalization across all languages and domains. We further propose attention sharing strategies to facilitate parameter sharing and specialization in multilingual and multi-domain sequence modeling. Our approach automatically learns shared and specialized attention heads for different languages and domains. Evaluated in various tasks including speech recognition, text-to-text and speech-to-text translation, the proposed attention sharing strategies consistently bring gains to sequence models built upon multi-head attention. For speech-to-text translation, our approach yields an average of $+2.0$ BLEU over 13 language directions in multilingual setting and $+2.0$ BLEU over 3 domains in multi-domain setting.

Hierarchical Skills for Efficient Exploration (NeurIPS, Dec 2021; Link)

Abstract:
In reinforcement learning, pre-trained low-level skills have the potential to greatly facilitate exploration. However, prior knowledge of the downstream task is required to strike the right balance between generality (fine-grained control) and specificity (faster learning) in skill design. In previous work on continuous control, the sensitivity of methods to this trade-off has not been addressed explicitly, as locomotion provides a suitable prior for navigation tasks, which have been of foremost interest. In this work, we analyze this trade-off for low-level policy pre-training with a new benchmark suite of diverse, sparse-reward tasks for bipedal robots. We alleviate the need for prior knowledge by proposing a hierarchical skill learning framework that acquires skills of varying complexity in an unsupervised manner. For utilization on downstream tasks, we present a three-layered hierarchical learning algorithm to automatically trade off between general and specific skills as required by the respective task. In our experiments, we show that our approach performs this trade-off effectively and achieves better results than current state-of-the-art methods for end-to-end hierarchical reinforcement learning and unsupervised skill discovery.

Interesting Object, Curious Agent: Learning Task-Agnostic Exploration (NeurIPS, Nov 2021; Link)

Abstract:
Common approaches for task-agnostic exploration learn tabula-rasa –the agent assumes isolated environments and no prior knowledge or experience. However, in the real world, agents learn in many environments and always come with prior experiences as they explore new ones. Exploration is a lifelong process. In this paper, we propose a paradigm change in the formulation and evaluation of task-agnostic exploration. In this setup, the agent first learns to explore across many environments without any extrinsic goal in a task-agnostic manner. Later on, the agent effectively transfers the learned exploration policy to better explore new environments when solving tasks. In this context, we evaluate several baseline exploration strategies and present a simple yet effective approach to learning taskagnostic exploration policies. Our key idea is that there are two components of exploration: (1) an agent-centric component encouraging exploration of unseen parts of the environment based on an agent's belief; (2) an environment-centric component encouraging exploration of inherently interesting objects. We show that our formulation is effective and provides the most consistent exploration across several training-testing environment pairs. We also introduce benchmarks and metrics for evaluating task-agnostic exploration strategies. The source code is available at https://github.com/sparisi/cbet/.

DOBF: A Deobfuscation Pre-Training Objective for Programming Languages (NeurIPS, Nov 2021; Link)

Abstract:
Recent advances in self-supervised learning have dramatically improved the state of the art on a wide variety of tasks. However, research in language model pre-training has mostly focused on natural languages, and it is unclear whether models like BERT and its variants provide the best pre-training when applied to other modalities, such as source code. In this paper, we introduce a new pre-training objective, DOBF, that leverages the structural aspect of programming languages and pre-trains a model to recover the original version of obfuscated source code. We show that models pre-trained with DOBF significantly outperform existing approaches on multiple downstream tasks, providing relative improvements of up to 12.2% in unsupervised code translation, and 5.3% in natural language code search. Incidentally, we found that our pre-trained model is able to deobfuscate fully obfuscated source files, and to suggest descriptive variable names.

MADE: Exploration via Maximizing Deviation from Explored Regions (NeurIPS, Nov 2021; Link)

Abstract:

In online reinforcement learning (RL), efficient exploration remains particularly challenging in high-dimensional environments with sparse rewards. In low-dimensional environments, where tabular parameterization is possible, count-based upper confidence bound (UCB) exploration methods achieve minimax near-optimal rates. However, it remains unclear how to efficiently implement UCB in realistic RL tasks that involve nonlinear function approximation. To address this, we propose a new exploration approach via maximizing the deviation of the occupancy of the next policy from the explored regions. We add this term as an adaptive regularizer to the standard RL objective to trade off between exploration and exploitation. We pair the new objective with a provably convergent algorithm, giving rise to a new intrinsic reward that adjusts existing bonuses. The proposed intrinsic reward is easy to implement and combine with other existing RL algorithms to conduct exploration. As a proof of concept, we evaluate the new intrinsic reward on tabular examples across a variety of model-based and model-free algorithms, showing improvements over count-only exploration strategies. When tested on navigation and locomotion tasks from MiniGrid and DeepMind Control Suite benchmarks, our approach significantly improves sample efficiency over state-of-the-art methods.

NovelD: A Simple yet Effective Exploration Criterion (NeurIPS, Nov 2021; Link)

Abstract:
Efficient exploration under sparse rewards remains a key challenge in deep reinforcement learning. Previous exploration methods (e.g., RND) have achieved strong results in multiple hard tasks. However, if there are multiple novel areas to explore, these methods often focus quickly on one without sufficiently trying others (like a depth-wise first search manner). In some scenarios (e.g., four corridor environment in Sec. 4.2), we observe they explore in one corridor for long and fail to cover all the states. On the other hand, in theoretical RL, with optimistic initialization and the inverse square root of visitation count as a bonus, it won't suffer from this and explores different novel regions alternatively (like a breadth-first search manner). In this paper, inspired by this, we propose a simple but effective criterion called NovelD by weighting every novel area approximately equally. Our algorithm is very simple but yet shows comparable performance or even outperforms multiple SOTA exploration methods in many hard exploration tasks. Specifically, NovelD solves all the static procedurally-generated tasks in Mini-Grid with just 120M environment steps, without any curriculum learning. In comparison, the previous SOTA only solves 50% of them. NovelD also achieves SOTA on multiple tasks in NetHack, a rogue-like game that contains more challenging procedurally-generated environments. In multiple Atari games (e.g., MonteZuma's Revenge, Venture, Gravitar), NovelD outperforms RND. We analyze NovelD thoroughly in MiniGrid and found that empirically it helps the agent explore the environment more uniformly with a focus on exploring beyond the boundary.

Decision Transformer: Reinforcement Learning via Sequence Modeling (NeurIPS, Oct 2021; Link)

Abstract:
We propose a hypothesis that effective policies can be learned from data without dynamic programming bootstrapping. To investigate this, we consider replacing traditional reinforcement learning (RL) algorithms -- which typically bootstrap against a learned value function -- with a simple sequence modeling objective. We train a transformer model on sequences of returns, states, and actions with an autoregressive prediction loss widely used in language modeling, reducing policy sampling to sequence generation. By training a transformer model using a supervised loss function, we can remove the need for dynamic programming bootstrapping, which is known to be unstable with function approximation. Furthermore, we can also leverage the simplicity, scalability, and long-range memory capabilities of transformers. Through experiments spanning a diverse set of offline RL benchmarks including Atari, OpenAI Gym, and Key-to-Door, we show that our Decision Transformer model can learn to generate diverse behaviors by conditioning on desired returns. In particular, our Decision Transformer, when conditioned with high desired returns, produces a policy that is competitive or better than state of the art model-free offline RL algorithms.

Visual Adversarial Imitation Learning using Variational Models (NeurIPS, Oct 2021; Link)

Abstract:
Reward function specification, which requires considerable human effort and iteration, remains a major impediment for learning behaviors through deep reinforcement learning. In contrast, providing visual demonstrations of desired behaviors often presents an easier and more natural way to teach agents. We consider a setting where an agent is provided a fixed dataset of visual demonstrations illustrating how to perform a task, and must learn to solve the task using the provided demonstrations and unsupervised environment interactions. This setting presents a number of challenges including representation learning for visual observations, sample complexity due to high dimensional spaces, and learning instability due to the lack of a fixed reward or learning signal. Towards addressing these challenges, we develop a variational model-based adversarial imitation learning (V-MAIL) algorithm. The model-based approach provides a strong signal for representation learning, enables sample efficiency, and improves the stability of adversarial training by enabling on-policy learning. Through experiments involving several vision-based locomotion and manipulation tasks, we find that V-MAIL learns successful visuomotor policies in a sample-efficient manner, has better stability compared to prior work, and also achieves higher asymptotic performance. We further find that by transferring the learned models, V-MAIL can learn new tasks from visual demonstrations without any additional environment interactions. All results including videos can be found online here.

Luna: Linear Unified Nested Attention (NeurIPS, Oct 2021; Link)

Abstract:
The quadratic computational and memory complexities of the Transformer's at-tention mechanism have limited its scalability for modeling long sequences. Inthis paper, we propose Luna, a linear unified nested attention mechanism thatapproximates softmax attention withtwo nested linear attention functions, yieldingonly linear (as opposed to quadratic) time and space complexity. As compared toa more traditional attention mechanism, Luna introduces an additional sequencewith a fixed length as input and an additional corresponding output, which allowsLuna to perform attention operation linearly, while also storing adequate contextualinformation. We perform extensive evaluations on three benchmarks of sequencemodeling tasks: long-context sequence modeling, neural machine translation andmasked language modeling for large-scale pretraining. Competitive or even betterexperimental results demonstrate both the effectiveness and efficiency of Lunacompared to a variety of strong baseline methods including the full-rank attentionand other efficient sparse and dense attention methods. The implementation of ourmodel is available here.

Unsupervised Speech Recognition (NeurIPS, Oct 2021; Link)

Abstract:
Despite rapid progress in the recent past, current speech recognition systems still require labeled training data which limits this technology to a small fraction of the languages spoken around the globe. This paper describes wav2vec-U, short for wav2vec Unsupervised, a method to train speech recognition models without any labeled data. We leverage self-supervised speech representations to segment unlabeled audio and learn a mapping from these representations to phonemes via adversarial training. The right representations are key to the success of our method. Compared to the best previous unsupervised work, wav2vec-U reduces the phone error rate on the TIMIT benchmark from 26.1 to 11.3. On the larger English Librispeech benchmark, wav2vec-U achieves a word error rate of 5.9 on test-other, rivaling some of the best published systems trained on 960 hours of labeled data from only two years ago. We also experiment on nine other languages, including low-resource languages such as Kyrgyz, Swahili and Tatar. The code will be open sourced.

Autoregressive Entity Retrieval (ICLR, May 2021; Link)

Abstract:
Entities are at the center of how we represent and aggregate knowledge. For instance, Encyclopedias such as Wikipedia are structured by entities (e.g., one per Wikipedia article). The ability to retrieve such entities given a query is fundamental for knowledge-intensive tasks such as entity linking and open-domain question answering. Current approaches can be understood as classifiers among atomic labels, one for each entity. Their weight vectors

are dense entity representations produced by encoding entity meta information such as their descriptions. This approach has several shortcomings: (i) context and entity affinity is mainly captured through a vector dot product, potentially missing fine-grained interactions; (ii) a large memory footprint is needed to store dense representations when considering large entity sets; (iii) an appropriately hard set of negative data has to be subsampled at training time. In this work, we propose GENRE, the first system that retrieves entities by generating their unique names, left to right, token-by-token in an autoregressive fashion. This mitigates the aforementioned technical issues since: (i) the autoregressive formulation directly captures relations between context and entity name, effectively cross encoding both; (ii) the memory footprint is greatly reduced because the parameters of our encoder-decoder architecture scale with vocabulary size, not entity count; (iii) the softmax loss is computed without subsampling negative data. We experiment with more than 20 datasets on entity disambiguation, end-to-end entity linking and document retrieval tasks, achieving new state-of-the-art or very competitive results while using a tiny fraction of the memory footprint of competing systems. Finally, we demonstrate that new entities can be added by simply specifying their names. Code and pre-trained models

Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation (NeurIPS, May 2021; Link)

Abstract:
This work addresses the challenge of hate speech detection in Internet memes, and attempts using visual information to automatically detect hate speech, unlike any previous work of our knowledge. Memes are pixel-based multimedia documents that contain photos or illustrations together with phrases which, when combined, usually adopt a funny meaning. However, hate memes are also used to spread hate through social networks, so their automatic detection would help reduce their harmful societal impact. Our results indicate that the model can learn to detect some of the memes, but that the task is far from being solved with this simple architecture. While previous work focuses on linguistic hate speech, our experiments indicate how the visual modality can be much more informative for hate speech detection than the linguistic one in memes. In our experiments, we built a dataset of 5,020 memes to train and evaluate a multi-layer perceptron over the visual and language representations, whether independently or fused.

Human-Level Performance in No-Press Diplomacy via Equilibrium Search (ICLR, May 2021; Link)

Abstract:
Prior AI breakthroughs in complex games have focused on either the purely adversarial or purely cooperative settings. In contrast, Diplomacy is a game of shifting alliances that involves both cooperation and competition. For this reason, Diplomacy has proven to be a formidable research challenge. In this paper we describe an agent for the no-press variant of

Diplomacy that combines supervised learning on human data with one-step lookahead search via regret minimization. Regret minimization techniques have been behind previous AI successes in adversarial games, most notably poker, but have not previously been shown to be successful in large-scale games involving cooperation. We show that our agent greatly exceeds the performance of past no-press Diplomacy bots, is unexploitable by expert humans, and ranks in the top 2% of human players when playing anonymous games on a popular Diplomacy website.

Answering Complex Open-domain Questions With Multi-hop Dense Retrieval (ICLR, Mar 2021; Link)

Abstract:
We propose a simple and efficient multi-hop dense retrieval approach for answering complex open-domain questions, which achieves state-of-the-art performance on two multi-hop datasets, HotpotQA and multi-evidence FEVER. Contrary to previous work, our method does not require access to any corpus-specific information, such as inter-document hyperlinks or human-annotated entity markers, and can be applied to any unstructured text corpus. Our system also yields a much better efficiency-accuracy trade-off, matching the best published accuracy on HotpotQA while being 10 times faster at inference time.

What they do when in doubt: a study of inductive biases in seq2seq learners (ICLR, May 2021; Link)

Abstract:
Sequence-to-sequence (seq2seq) learners are widely used, but we still have only limited knowledge about what inductive biases shape the way they generalize. We address that by investigating how popular seq2seq learners generalize in tasks that have high ambiguity in the training data. We use four new tasks to study learners' preferences for memorization, arithmetic, hierarchical, and compositional reasoning. Further, we connect to Solomonoff's theory of induction and propose to use description length as a principled and sensitive measure of inductive biases. In our experimental study, we find that LSTM-based learners can learn to perform counting, addition, and multiplication by a constant from a single training example. Furthermore, Transformer and LSTM-based learners show a bias toward the hierarchical induction over the linear one, while CNN-based learners prefer the opposite. The latter also show a bias toward a compositional generalization over memorization. Finally, across all our experiments, description length proved to be a sensitive measure of inductive biases.

WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia (EACL, Apr 2021; Link)

Abstract:
We present an approach based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages, including several dialects or low-resource languages. We systematically consider all possible language pairs. In total, we are able to extract 135M parallel sentences for 1620 different language pairs, out of which only 34M are aligned with English. This corpus is freely available. To get an indication on the quality of the extracted bitexts, we train neural MT baseline systems on the mined data only for 1886 languages pairs, and evaluate them on the TED corpus, achieving strong BLEU scores for many language pairs. The WikiMatrix bitexts seem to be particularly interesting to train MT systems between distant languages without the need to pivot through English.

Semantic Audio-Visual Navigation (CVPR, Apr 2021; Link)

Abstract:
Recent work on audio-visual navigation assumes a constantly-sounding target and restricts the role of audio to signaling the target's position. We introduce semantic audio-visual navigation, where objects in the environment make sounds consistent with their semantic meaning (e.g., toilet flushing, door creaking) and acoustic events are sporadic or short in duration. We propose a transformer-based model to tackle this new semantic AudioGoal task, incorporating an inferred goal descriptor that captures both spatial and semantic properties of the target. Our model's persistent multimodal memory enables it to reach the goal even long after the acoustic event stops. In support of the new task, we also expand the SoundSpaces audio simulations to provide semantically grounded sounds for an array of objects in Matterport3D. Our method strongly outperforms existing audio-visual navigation methods by learning to associate semantic, acoustic, and visual cues. Project page.

Bi-directional Domain Adaptation for Sim2Real Transfer of Embodied Navigation Agents (RA-L, Apr 2021; Link)

Abstract:
Deep reinforcement learning models are notoriously data hungry, yet real-world data is expensive and time consuming to obtain. The solution that many have turned to is to use simulation for training before deploying the robot in a real environment. Simulation offers the ability to train large numbers of robots in parallel, and offers an abundance of data. However, no simulation is perfect, and robots trained solely in simulation fail to generalize to the real-world, resulting in a "sim-vs-real gap". How can we overcome the trade-off between the abundance of less accurate, artificial data from simulators and the scarcity of reliable,

real-world data? In this paper, we propose Bi-directional Domain Adaptation (BDA), a novel approach to bridge the sim-vs-real gap in both directions -- real2sim to bridge the visual domain gap, and sim2real to bridge the dynamics domain gap. We demonstrate the benefits of BDA on the task of PointGoal Navigation. BDA with only 5k real-world (state, action, next-state) samples matches the performance of a policy fine-tuned with ~600k samples, resulting in a speed-up of ~120×.

MixKD: Towards Efficient Distillation of Large-scale Language Models (ICLR, Mar 2021; Link)

Abstract:
Large-scale language models have recently demonstrated impressive empirical performance. Nevertheless, the improved results are attained at the price of bigger models, more power consumption, and slower inference, which hinder their applicability to low-resource (both memory and computation) platforms. Knowledge distillation (KD) has been demonstrated as an effective framework for compressing such big models. However, large-scale neural network systems are prone to memorize training instances, and thus tend to make inconsistent predictions when the data distribution is altered slightly. Moreover, the student model has few opportunities to request useful information from the teacher model when there is limited task-specific data available. To address these issues, we propose MixKD, a data-agnostic distillation framework that leverages mixup, a simple yet efficient data augmentation approach, to endow the resulting model with stronger generalization ability. Concretely, in addition to the original training examples, the student model is encouraged to mimic the teacher's behavior on the linear interpolation of example pairs as well. We prove from a theoretical perspective that under reasonable conditions MixKD gives rise to a smaller gap between the generalization error and the empirical error. To verify its effectiveness, we conduct experiments on the GLUE benchmark, where MixKD consistently leads to significant gains over the standard KD training, and outperforms several competitive baselines. Experiments under a limited-data setting and ablation studies further demonstrate the advantages of the proposed approach.

On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning (AISTATS, Mar 2021; Link)

Abstract:
Model-based Reinforcement Learning (MBRL) is a promising framework for learning control in a data-efficient manner. MBRL algorithms can be fairly complex due to the separate dynamics modeling and the subsequent planning algorithm, and as a result, they often possess tens of hyperparameters and architectural choices. For this reason, MBRL typically requires significant human expertise before it can be applied to new problems and domains. To alleviate this problem, we propose to use automatic hyperparameter optimization (HPO). We demonstrate that this problem can be tackled effectively with automated HPO, which we

demonstrate to yield significantly improved performance compared to human experts. In addition, we show that tuning of several MBRL hyperparameters dynamically, i.e. during the training itself, further improves the performance compared to using static hyperparameters which are kept fixed for the whole training. Finally, our experiments provide valuable insights into the effects of several hyperparameters, such as plan horizon or learning rate and their influence on the stability of training and resulting rewards.

Reinforcement Learning-based Product Delivery Frequency Control (AAAI, Dec 2020; Link)

Abstract:
Frequency control is an important problem in modern recommender systems. It dictates the delivery frequency of recommendations to maintain product quality and efficiency. For example, the frequency of delivering promotional notifications impacts daily metrics as well as the infrastructure resource consumption (e.g. CPU and memory usage). There remain open questions on what objective we should optimize to represent business values in the long term best, and how we should balance between daily metrics and resource consumption in a dynamically fluctuating environment. We propose a personalized methodology for the frequency control problem, which combines long-term value optimization using reinforcement learning (RL) with a robust volume control technique we termed "Effective Factor". We demonstrate statistically significant improvement in daily metrics and resource efficiency by our method in several notification applications at a scale of billions of users. To our best knowledge, our study represents the first deep RL application on the frequency control problem at such an industrial scale.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (NeurIPS, Dec 2020; Link)

Abstract:
Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same

retrieved passages across the whole generated sequence, the other can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

Resource Constrained Dialog Policy Learning via Differentiable Inductive Logic Programming (COLING, Dec 2020; Link)

Abstract:
Motivated by the needs of resource constrained dialog policy learning, we introduce dialog policy via differentiable inductive logic (DILOG). We explore the tasks of one-shot learning and zero-shot domain transfer with DILOG on SimDial and MultiWoZ. Using a single representative dialog from the restaurant domain, we train DILOG on the SimDial dataset and obtain 99+% in-domain test accuracy. We also show that the trained DILOG zero-shot transfers to all other domains with 99+% accuracy, proving the suitability of DILOG to slot-filling dialogs. We further extend our study to the MultiWoZ dataset achieving 90+% inform and success metrics. We also observe that these metrics are not capturing some of the shortcomings of DILOG in terms of false positives, prompting us to measure an auxiliary Action F1 score. We show that DILOG is 100x more data efficient than state-of-the-art neural approaches on MultiWoZ while achieving similar performance metrics. We conclude with a discussion on the strengths and weaknesses of DILOG.

Situated and Interactive Multimodal Conversations (COLING, Dec 2020; Link)

Abstract:
Next generation virtual assistants are envisioned to handle multimodal inputs (e.g., vision, memories of previous interactions, and the user's utterances), and perform multimodal actions (e.g., displaying a route while generating the system's utterance). We introduce Situated Interactive MultiModal Conversations (SIMMC) as a new direction aimed at training agents that take multimodal actions grounded in a co-evolving multimodal input context in addition to the dialog history. We provide two SIMMC datasets totalling ~13K human-human dialogs (~169K utterances) collected using a multimodal Wizard-of-Oz (WoZ) setup, on two shopping domains: (a) furniture – grounded in a shared virtual environment; and (b) fashion – grounded in an evolving set of images. Datasets include multimodal context of the items appearing in each scene, and contextual NLU, NLG and coreference annotations using a novel and unified framework of SIMMC conversational acts for both user and assistant utterances.

Finally, we present several tasks within SIMMC as objective evaluation protocols, such as structural API prediction, response generation, and dialog state tracking. We benchmark a collection of existing models on these SIMMC tasks as strong baselines, and demonstrate rich multimodal conversational interactions. Our data, annotations, and models are publicly available.

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (NeurIPS, Dec 2020; Link)

Abstract:
We show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.

Pre-training via Paraphrasing (NeurIPS, Dec 2020; Link)

Abstract:
We introduce MARGE, a pre-trained sequence-to-sequence model learned with an unsupervised multi-lingual multi-document paraphrasing objective. MARGE provides an alternative to the dominant masked language modeling paradigm, where we self-supervise the reconstruction of target text by retrieving a set of related texts (in many languages) and conditioning on them to maximize the likelihood of generating the original. We show it is possible to jointly learn to do retrieval and reconstruction, given only a random initialization. The objective noisily captures aspects of paraphrase, translation, multi-document summarization, and information retrieval, allowing for strong zero-shot performance on several tasks. For example, with no additional task-specific training we achieve BLEU scores of up to 35.8 for document translation. We further show that fine-tuning gives strong performance on a range of discriminative and generative tasks in many languages, making MARGE the most generally applicable pre-training method to date.

Massively Multilingual Document Alignment with Cross-lingual Sentence-Mover's Distance (AACL, Dec 2020; Link)

Abstract:
Document alignment aims to identify pairs of documents in two distinct languages that are of comparable content or translations of each other. Such aligned data can be used for a variety of NLP tasks from training cross-lingual representations to mining parallel data for machine translation. In this paper we develop an unsupervised scoring function that leverages cross-lingual sentence embeddings to compute the semantic distance between documents in different languages. These semantic distances are then used to guide a document alignment algorithm to properly pair cross-lingual web documents across a variety of low, mid, and high-resource language pairs. Recognizing that our proposed scoring function and other state of the art methods are computationally intractable for long web documents, we utilize a more tractable greedy algorithm that performs comparably. We experimentally demonstrate that our distance metric performs better alignment than current baselines outperforming them by 7% on high-resource language pairs, 15% on mid-resource language pairs, and 22% on low-resource language pairs.

Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning (ArXiV, Nov 2020; Link)

Abstract:
Accurate reporting of energy and carbon usage is essential for understanding the potential climate impacts of machine learning research. We introduce a framework that makes this easier by providing a simple interface for tracking realtime energy consumption and carbon emissions, as well as generating standardized online appendices. Utilizing this framework, we create a leaderboard for energy efficient reinforcement learning algorithms to incentivize responsible research in this area as an example for other areas of machine learning. Finally, based on case studies using our framework, we propose strategies for mitigation of carbon emissions and reduction of energy consumption. By making accounting easier, we hope to further the sustainable development of machine learning experiments and spur more research into energy efficient algorithms.

Dense Passage Retrieval for Open-Domain Question Answering (EMNLP, Nov 2020; Link)

Abstract:
Open-domain question answering relies on efficient passage retrieval to select candidate contexts, where traditional sparse vector space models, such as TF-IDF or BM25, are the de facto method. In this work, we show that retrieval can be practically implemented using dense representations alone, where embeddings are learned from a small number of

questions and passages by a simple dual-encoder framework. When evaluated on a wide range of open-domain QA datasets, our dense retriever outperforms a strong Lucene-BM25 system greatly by 9%-19% absolute in terms of top-20 passage retrieval accuracy, and helps our end-to-end QA system establish new state-of-the-art on multiple open-domain QA benchmarks.

Scalable Zero-shot Entity Linking with Dense Entity Retrieval (EMNLP, Nov 2020; Link)

Abstract:
This paper introduces a conceptually simple, scalable, and highly effective BERT-based entity linking model, along with an extensive evaluation of its accuracy-speed trade-off. We present a two-stage zero-shot linking algo- rithm, where each entity is defined only by a short textual description. The first stage does retrieval in a dense space defined by a bi-encoder that independently embeds the mention context and the entity descriptions. Each candidate is then re-ranked with a cross- encoder, that concatenates the mention and en- tity text. Experiments demonstrate that this approach is state of the art on recent zero- shot benchmarks (6 point absolute gains) and also on more established non-zero-shot evaluations (e.g. TACKBP-2010), despite its rel- ative simplicity (e.g. no explicit entity embeddings or manually engineered mention ta- bles). We also show that bi-encoder link- ing is very fast with nearest neighbour search (e.g. linking with 5.9 million candidates in 2 milliseconds), and that much of the ac- curacy gain from the more expensive cross- encoder can be transferred to the bi-encoder via knowledge distillation. Our code and models are available at https://github. com/facebookresearch/BLINK.

CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs (EMNLP, Nov 2020; Link)

Abstract:
Cross-lingual document alignment aims to identify pairs of documents in two distinct languages that are of comparable content or translations of each other. In this paper, we exploit the signals embedded in URLs to label web documents at scale with an average precision of 94.5% across different language pairs. We mine sixty-eight snapshots of the Common Crawl corpus and identify web document pairs that are translations of each other. We release a new web dataset consisting of over 392 million URL pairs from Common Crawl covering documents in 8144 language pairs of which 137 pairs include English. In addition to curating this massive dataset, we introduce baseline methods that leverage cross-lingual representations to identify aligned documents based on their textual content. Finally, we demonstrate the value of this parallel documents dataset through a downstream task of mining parallel sentences and measuring the quality of machine translations from models trained on this mined data. Our objective in releasing this dataset is to foster new research in cross-lingual NLP across a variety of low, medium, and high-resource languages.

Sound Natural: Content Rephrasing in Dialog Systems (EMNLP, Nov 2020; Link)

Abstract:
We introduce a new task of rephrasing for a more natural virtual assistant. Currently, virtual assistants work in the paradigm of intent slot tagging and the slot values are directly passed as-is to the execution engine. However, this setup fails in some scenarios such as messaging when the query given by the user needs to be changed before repeating it or sending it to another user. For example, for queries like 'ask my wife if she can pick up the kids' or 'remind me to take my pills', we need to rephrase the content to 'can you pick up the kids' and 'take your pills'. In this paper, we study the problem of rephrasing with messaging as a use case and release a dataset of 3000 pairs of original query and rephrased query. We show that BART, a pre-trained transformers-based masked language model with auto-regressive decoding, is a strong baseline for the task, and show improvements by adding a copy-pointer and copy loss to it. We analyze different tradeoffs of BART-based and LSTM-based seq2seq models, and propose a distilled LSTM-based seq2seq as the best practical model.

Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art (EMNLP, Nov 2020; Link)

Abstract:
A large array of pretrained models are available to the biomedical NLP (BioNLP) community. Finding the best model for a particular task can be difficult and time-consuming. For many applications in the biomedical and clinical domains, it is crucial that models can be built quickly and are highly accurate. We present a large-scale study across 18 established biomedical and clinical NLP tasks to determine which of several popular open-source biomedical and clinical NLP models work well in different settings. Furthermore, we apply recent advances in pretraining to train new biomedical language models, and carefully investigate the effect of various design choices on downstream performance. Our best models perform well in all of our benchmarks, and set new State-of-the-Art in 9 tasks. We release these models in the hope that they can help the community to speed up and increase the accuracy of BioNLP and text mining applications.

Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English (COLING, Oct 2020; Link)

Abstract:
We conduct in this work an evaluation study comparing offline and online neural machine translation architectures. Two sequence-to-sequence models: convolutional Pervasive

Attention (Elbayad et al., 2018) and attention-based Transformer (Vaswani et al., 2017) are considered. We investigate, for both architectures, the impact of online decoding constraints on the translation quality through a carefully designed human evaluation on English-German and German-English language pairs, the latter being particularly sensitive to latency constraints. The evaluation results allow us to identify the strengths and shortcomings of each model when we shift to the online setup.

Recipes for Safety in Open-domain Chatbots (ArXiV, Oct 2020; Link)

Abstract:
Models trained on large unlabeled corpora of human interactions will learn patterns and mimic behaviors therein, which include offensive or otherwise toxic behavior and unwanted biases. We investigate a variety of methods to mitigate these issues in the context of open-domain generative dialogue models. We introduce a new human-and-model-in-the-loop framework for both training safer models and for evaluating them, as well as a novel method to distill safety considerations inside generative models without the use of an external classifier at deployment time. We conduct experiments comparing these methods and find our new techniques are (i) safer than existing models as measured by automatic and human evaluations while (ii) maintaining usability metrics such as engagingness relative to the state of the art. We then discuss the limitations of this work by analyzing failure cases of our models.

Neural Database Operator Model (WeCNLP, Oct 2020; Link)

Abstract:
Our goal is to answer queries over facts stored in a text memory. The key challenge in NeuralDBs(Thorne et al., 2020), compared to open-book NLP such as question answering (Rajpurkar et al., 2016,inter alia), is that possibly thousands of facts must be aggregated to provide a single answer, without direct supervision. The challenges represented in NeuralDBs are important for both the NLP and database communities alike: discrete reasoning over text (Dua et al., 2019), retriever-based QA(Dunn et al., 2017) and multi-hop QA (Welbl et al.,2018; Yang et al., 2018) are common components.

Large-scale Pretraining for Visual Dialog:A Simple State-of-the-Art Baseline (ECCV, Jul 2020; Link)

Abstract:
Prior work in visual dialog has focused on training deep neural models on VisDial [1] in isolation. Instead, we present an approach to leverage pretraining on related vision-language

datasets before transferring to visual dialog. We adapt the recently proposed ViLBERT model [2] for multi-turn visually-grounded conversations. Our model is pretrained on the Conceptual Captions [3] and Visual Question Answering [4] datasets, and finetuned on VisDial. Our best single model outperforms prior published work by >1% absolute on NDCG and MRR. Next, we find that additional finetuning using “dense” annotations in VisDial leads to even higher NDCG – more than 10% over our base model – but hurts MRR – more than 17% below our base model! This highlights a trade-off between the two primary metrics – NDCG and MRR – which we find is due to dense annotations not correlating well with the original ground-truth answers to questions.

Spatially Aware Multimodal Transformers for TextVQA (ECCV, Jul 2020; Link)

Abstract:
Textual cues are essential for everyday tasks like buying groceries and using public transport. To develop this assistive technology, we study the TextVQA task, i.e., reasoning about text in images to answer a question. Existing approaches are limited in their use of spatial relations and rely on fully-connected transformer-based architectures to implicitly learn the spatial structure of a scene. In contrast, we propose a novel spatially aware self-attention layer such that each visual entity only looks at neighboring entities defined by a spatial graph. Further, each head in our multi-head self-attention layer focuses on a different subset of relations. Our approach has two advantages: (1) each head considers local context instead of dispersing the attention amongst all visual entities; (2) we avoid learning redundant features. We show that our model improves the absolute accuracy of current state-of-the-art methods on TextVQA by 2.2% overall over an improved baseline, and 4.62% on questions that involve spatial reasoning and can be answered correctly using OCR tokens. Similarly on ST-VQA, we improve the absolute accuracy by 4.2%. We further show that spatially aware self-attention improves visual grounding.

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (ACL, Jul 2020; Link)

Abstract:
We present BART, a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and other recent pre-training schemes. We evaluate a number of noising approaches, finding the best perfor-mance by both randomly shuffling the order of sentences and using a novel in-filling

scheme,where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa on GLUE and SQuAD, and achieves new state-of-the-art results on a range of abstractive di-alogue, question answering, and summarization tasks, with gains of up to 3.5 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pretraining. We also replicate other pretraining schemes within the BART framework, to understand their effect on end-task performance.

On The Evaluation of Machine Translation Systems Trained With Back-Translation (ACL, Jul 2020; Link)

Abstract:
Back-translation is a widely used data augmentation technique which leverages target monolingual data. However, its effectiveness has been challenged since automatic metrics such as BLEU only show significant improvements for test examples where the source itself is a translation, or translationese. This is believed to be due to translationese inputs better matching the back-translated training data. In this work, we show that this conjecture is not empirically supported and that backtranslation improves translation quality of both naturally occurring text as well as translationese according to professional human translators. We provide empirical evidence to support the view that back-translation is preferred by humans because it produces more fluent outputs. BLEU cannot capture human preferences because references are translationese when source sentences are natural text. We recommend complementing BLEU with a language model score to measure fluency.

Information-Theoretic Probing for Linguistic Structure (ACL, Jun 2020; Link)

Abstract:
The success of neural networks on a diverse set of NLP tasks has led researchers to question how much these networks actually "know" about natural language. Probes are a natural way of assessing this. When probing, a researcher chooses a linguistic task and trains a supervised model to predict annotations in that linguistic task from the network's learned representations. If the probe does well, the researcher may conclude that the representations encode knowledge related to the task. A commonly held belief is that using simpler models as probes is better; the logic is that simpler models will identify linguistic structure, but not learn the task itself. We propose an information-theoretic operationalization of probing as estimating mutual information that contradicts this received wisdom: one should always select the highest performing probe one can, even if it is more complex, since it will result in a tighter estimate, and thus reveal more of the linguistic information inherent in the representation. The experimental portion of our paper focuses on empirically estimating the mutual information between a linguistic property and BERT, comparing these estimates

to several baselines. We evaluate on a set of ten typologically diverse languages often underrepresented in NLP research—plus English—totaling eleven languages.

Language Models as Fact Checkers? (ACL, Jun 2020; Link)

Abstract:
Recent work has suggested that language models (LMs) store both common-sense and factual knowledge learned from pre-training data. In this paper, we leverage this implicit knowledge to create an effective end-to-end fact checker using a solely a language model, without any external knowledge or explicit retrieval components. While previous work on extracting knowledge from LMs have focused on the task of open-domain question answering, to the best of our knowledge, this is the first work to examine the use of language models as fact checkers. In a closed-book setting, we show that our zero-shot LM approach outperforms a random baseline on the standard FEVER task, and that our finetuned LM compares favorably with standard baselines. Though we do not ultimately outperform methods which use explicit knowledge bases, we believe our exploration shows that this method is viable and has much room for exploration

Learning an Unreferenced Metric for Online Dialogue Evaluation (ACL, Jun 2020; Link)

Abstract:
Evaluating the quality of a dialogue interaction between two agents is a difficult task, especially in open-domain chit-chat style dialogue. There have been recent efforts to develop automatic dialogue evaluation metrics, but most of them do not generalize to unseen datasets and/or need a human-generated reference response during inference, making it infeasible for online evaluation. Here, we propose an unreferenced automated evaluation metric that uses large pre-trained language models to extract latent representations of utterances, and leverages the temporal transitions that exist between them. We show that our model achieves higher correlation with human annotations in an online setting, while not requiring true responses for comparison during inference.

Joint Modelling of Emotion and Abusive Language Detection (ACL, Jun 2020; Link)

Abstract:
The rise of online communication platforms has been accompanied by some undesirable effects, such as the proliferation of aggressive and abusive behaviour online. Aiming to tackle this problem, the natural language processing (NLP) community has experimented

with a range of techniques for abuse detection. While achieving substantial success, these methods have so far only focused on modelling the linguistic properties of the comments and the online communities of users, disregarding the emotional state of the users and how this might affect their language. The latter is, however, inextricably linked to abusive behaviour. In this paper, we present the first joint model of emotion and abusive language detection, experimenting in a multi-task learning framework that allows one task to inform the other. Our results demonstrate that incorporating affective features leads to significant improvements in abuse detection performance across datasets.

Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training (ACL, Jun 2020; Link)

Abstract:
Generative dialogue models currently suffer from a number of problems which standard maximum likelihood training does not address. They tend to produce generations that (i) rely too much on copying from the context, (ii) contain repetitions within utterances, (iii) overuse frequent words, and (iv) at a deeper level, contain logical flaws. In this work we show how all of these problems can be addressed by extending the recently introduced unlikelihood loss (Welleck et al., 2019a) to these cases. We show that appropriate loss functions which regularize generated outputs to match human distributions are effective for the first three issues. For the last important general issue, we show applying unlikelihood to collected data of what a model should not do is effective for improving logical consistency, potentially paving the way to generative models with greater reasoning ability. We demonstrate the efficacy of our approach across several dialogue tasks.

CamemBERT: a Tasty French Language Model (ACL, Jun 2020; Link)

Abstract:
Pretrained language models are now ubiquitous in Natural Language Processing. Despite their success, most available models have either been trained on English data or on the concatenation of data in multiple languages. This makes practical use of such models—in all languages except English—very limited. In this paper, we investigate the feasibility of training monolingual Transformer based language models for other languages, taking French as an example and evaluating our language models on part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks. We show that the use of web crawled data is preferable to the use of Wikipedia data. More surprisingly, we show that a relatively small web crawled dataset (4GB) leads to results that are as good as those obtained using larger datasets (130+GB). Our best performing model CamemBERT reaches or improves the state of the art in all four downstream tasks.

ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations (ACl, Jun 2020; Link)

Abstract:
In order to simplify a sentence, human editors perform multiple rewriting transformations: they split it into several shorter sentences, paraphrase words (i.e. replacing complex words or phrases by simpler synonyms), reorder components, and/or delete information deemed unnecessary. Despite these varied range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on a single transformation, such as lexical paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more realistic settings. To alleviate this limitation, this paper introduces ASSET, a new dataset for assessing sentence simplification in English. ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we show that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivate the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable when multiple simplification transformations are performed.

TaBert: Pretraining for Joint Understanding of Textual and Tabular Data (ACL, May 2020; Link)

Abstract:
Recent years have witnessed the burgeoning of pretrained language models (LMs) for text-based natural language (NL) understanding tasks. Such models are typically trained on free-form NL text, hence may not be suitable for tasks like semantic parsing over structured data, which require reasoning over both free-form NL questions and structured tabular data (e.g., database tables). In this paper we present TABERT, a pretrained LM that jointly learns representations for NL sentences and (semi-)structured tables. TaBert is trained on a large corpus of 26 million tables and their English contexts. In experiments, neural semantic parsers using TABERT as feature representation layers achieve new best results on the challenging weakly-supervised semantic parsing benchmark WikiTableQuestions, while performing competitively on the text-to-SQL dataset Spider.

Effectiveness of self-supervised pre-training for ASR (ICASSP, Apr 2020; Link)

Abstract:

We compare self-supervised representation learning algorithms which either explicitly quantize the audio data or learn representations without quantization. We find the former to be more accurate since it builds a good vocabulary of the data through vq-wav2vec [1] self-supervision approach to enable learning of effective representations in subsequent BERT training. Different to previous work, we directly fine-tune the pre-trained BERT models on transcribed speech using a Connectionist Temporal Classification (CTC) loss instead of feeding the representations into a task-specific model. We also propose a BERT-style model learning directly from the continuous audio data and compare pre-training on raw audio to spectral features. Fine-tuning a BERT model on 10 hour of labeled Librispeech data with a vq-wav2vec vocabulary is almost as good as the best known reported system trained on 100 hours of labeled data on test-clean, while achieving a 25% WER reduction on test-other. When using only 10 minutes of labeled data, WER is 25.2 on test-other and 16.3 on test-clean. This demonstrates that self-supervision can enable speech recognition systems trained on a near-zero amount of transcribed data.

The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents (ACL, Apr 2020; Link)

Abstract:
We introduce dodecaDialogue: a set of 12 tasks that measures if a conversational agent can communicate engagingly with personality and empathy, ask questions, answer questions by utilizing knowledge resources, discuss topics and situations, and perceive and converse about images. By multi-tasking on such a broad large-scale set of data, we hope to both move towards and measure progress in producing a single unified agent that can perceive, reason and converse with humans in an open-domain setting. We show that such multi-tasking improves over a BERT pre-trained baseline, largely due to multi-tasking with very large dialogue datasets in a similar domain, and that the multi-tasking in general provides gains to both text and image-based tasks using several metrics in both the fine-tune and task transfer settings. We obtain state-of-the-art results on many of the tasks, providing a strong baseline for this challenge.

Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills (ACL, Apr 2020; Link)

Abstract:
Being engaging, knowledgeable, and empathetic are all desirable general qualities in a conversational agent. Previous work has introduced tasks and datasets that aim to help agents to learn those qualities in isolation and gauge how well they can express them. But rather than being specialized in one single quality, a good open-domain conversational agent should be able to seamlessly blend them all into one cohesive conversational flow. In this work, we investigate several ways to combine models trained towards isolated capabilities, ranging from simple model aggregation schemes that require minimal additional training, to

various forms of multi-task training that encompass several skills at all training stages. We further propose a new dataset, BlendedSkillTalk, to analyze how these capabilities would mesh together in a natural conversation, and compare the performance of different architectures and training schemes. Our experiments show that multi-tasking over several tasks that focus on particular capabilities results in better blended conversation performance compared to models trained on a single skill, and that both unified or two-stage approaches perform well if they are constructed to avoid unwanted bias in skill selection or are fine-tuned on our new task.

RTFM: Generalizing to Novel Environment via Reading (ICLR, Mar 2020; Link)

Abstract:
Obtaining policies that can generalise to new environments in reinforcement learning is challenging. In this work, we demonstrate that language understanding via a reading policy learner is a promising vehicle for generalisation to new environments. We propose a grounded policy learning problem, Read to Fight Monsters (RTFM), in which the agent must jointly reason over a language goal, relevant dynamics described in a document, and environment observations. We procedurally generate environment dynamics and corresponding language descriptions of the dynamics, such that agents must read to understand new environment dynamics instead of memorising any particular information. In addition, we propose txt2π, a model that captures three-way interactions between the goal, document, and observations. On RTFM, txt2π generalises to new environments with dynamics not seen during training via reading. Furthermore, our model outperforms baselines such as FiLM and language-conditioned CNNs on RTFM. Through curriculum learning, txt2π produces policies that excel on complex RTFM tasks requiring several reasoning and coreference steps.

Compositional generalization through meta sequence-to-sequence learning (NeurIPS, Dec 2019; Link)

Abstract:
People can learn a new concept and use it compositionally, understanding how to “blicket twice” after learning how to “blicket.” In contrast, powerful sequence-to-sequence (seq2seq) neural networks fail such tests of compositionality, especially when composing new concepts together with existing concepts. In this paper, I show how memory-augmented neural networks can be trained to generalize compositionally through meta seq2seq learning. In this approach, models train on a series of seq2seq problems to acquire the compositional skills needed to solve new seq2seq problems. Meta seq2seq learning solves several of the SCAN tests for compositional learning and can learn to apply implicit rules to variables.

Hyperbolic Graph Neural Networks (NeurIPS, Dec 2019; Link)

Abstract:
Learning from graph-structured data is an important task in machine learning and artificial intelligence, for which Graph Neural Networks (GNNs) have shown great promise. Motivated by recent advances in geometric representation learning, we propose a novel GNN architecture for learning representations on Riemannian manifolds with differentiable exponential and logarithmic maps. We develop a scalable algorithm for modeling the structural properties of graphs, comparing Euclidean and hyperbolic geometry. In our experiments, we show that hyperbolic GNNs can lead to substantial improvements on various benchmark datasets.

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems (NeurIPS, Dec 2019; Link)

Abstract:
In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced a little over one year ago, offers a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research. In this paper we present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, a software toolkit, and a public leaderboard. SuperGLUE is available here.

RUBi: Reducing Unimodal Biases for Visual Question Answering (NeurIPS, Nov 2019; Link)

Abstract:
Visual Question Answering (VQA) is the task of answering questions about an image. Some VQA models often exploit unimodal biases to provide the correct answer without using the image information. As a result, they suffer from a huge drop in performance when evaluated on data outside their training set distribution. This critical issue makes them unsuitable for real-world settings. We propose RUBi, a new learning strategy to reduce biases in any VQA model. It reduces the importance of the most biased examples, i.e. examples that can be correctly classified without looking at the image. It implicitly forces the VQA model to use the two input modalities instead of relying on statistical regularities between the question and the answer. We leverage a question-only model that captures the language biases by identifying when these unwanted regularities are used. It prevents the base VQA model from learning them by influencing its predictions. This leads to dynamically adjusting the loss in order to compensate for biases. We validate our contributions by surpassing the current

state-of-the-art results on VQA-CP v2. This dataset is specifically designed to assess the robustness of VQA models when exposed to different question biases at test time than what was seen during training. Our code is available: github.com/cdancette/rubi.bootstrap.pytorch

Memory Grounded Conversational Reasoning (EMNLP, Nov 2019; Link)

Abstract:
We demonstrate a conversational system which engages the user through a multi-modal, multi-turn dialog over the user's memories. The system can perform QA over memories by responding to user queries to recall specific attributes and associated media (e.g. photos) of past episodic memories. The system can also make proactive suggestions to surface related events or facts from past memories to make conversations more engaging and natural. To implement such a system, we collect a new corpus of memory grounded conversations, which comprises human-to-human role-playing dialogs given synthetic memory graphs with simulated attributes. Our proof-of-concept system operates on these synthetic memory graphs, however it can be trained and applied to real-world user memory data (e.g. photo albums, etc.) We present the architecture of the proposed conversational system, and example queries that the system supports.

Language Models as Knowledge Bases? (EMNLP, Nov 2019; Link)

Abstract:
Recent progress in pretraining language models on large textual corpora led to a surge of improvements for downstream NLP tasks. Whilst learning linguistic knowledge, these models may also be storing relational knowledge present in the training data, and may be able to answer queries structured as "fill-in-the-blank" cloze statements. Language models have many advantages over structured knowledge bases: they require no schema engineering, allow practitioners to query about an open class of relations, are easy to extend to more data, and require no human supervision to train. We present an in-depth analysis of the relational knowledge already present (without fine-tuning) in a wide range of state-of-the-art pretrained language models. We find that (i) without fine-tuning, BERT contains relational knowledge competitive with traditional NLP methods that have some access to oracle knowledge, (ii) BERT also does remarkably well on open-domain question answering against a supervised baseline, and (iii) certain types of factual knowledge are learned much more readily than others by standard language model pretraining approaches. The surprisingly strong ability of these models to recall factual knowledge without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems. The code to reproduce our analysis is available here.

VizSeq: A Visual Analysis Toolkit for Text Generation Tasks (EMNLP, Nov 2019; Link)

Abstract:
Automatic evaluation of text generation tasks (e.g. machine translation, text summarization, image captioning and video description) usually relies heavily on task-specific metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). They, however, are abstract numbers and are not perfectly aligned with human assessment. This suggests inspecting detailed examples as a complement to identify system error patterns. In this paper, we present VizSeq, a visual analysis toolkit for instance-level and corpus-level system evaluation on a wide variety of text generation tasks. It supports multimodal sources and multiple text references, providing visualization in Jupyter notebook or a web app interface. It can be used locally or deployed onto public servers for centralized data hosting and benchmarking. It covers most common n-gram based metrics accelerated with multiprocessing, and also provides latest embedding-based metrics such as BERTScore (Zhang et al., 2019).

Finding Generalizable Evidence by Learning to Convince Q&A Models (EMNLP, Nov 2019; Link)

Abstract:
We propose a system that finds the strongest supporting evidence for a given answer to a question, using passage-based question-answering (QA) as a testbed. We train evidence agents to select the passage sentences that most convince a pretrained QA model of a given answer, if the QA model received those sentences instead of the full passage. Rather than finding evidence that convinces one model alone, we find that agents select evidence that generalizes; agent-chosen evidence increases the plausibility of the supported answer, as judged by other QA models and humans. Given its general nature, this approach improves QA in a robust manner: using agent-selected evidence (i) humans can correctly answer questions with only ~20% of the full passage and (ii) QA models can generalize to longer passages and harder questions.

Improving Generative Visual Dialog by Answering Diverse Questions (EMNLP, Nov 2019; Link)

Abstract:
Prior work on training generative Visual Dialog models with reinforcement learning (Das et al., 2017b) has explored a Q-BOT-A-BOT image-guessing game and shown that this 'self-talk' approach can lead to improved performance at the downstream dialog-conditioned image-guessing task. However, this improvement saturates and starts degrading after a few rounds of interaction, and does not lead to a better Visual Dialog model. We find that this is

due in part to repeated interactions between Q-BOT and A-BOT during self-talk, which are not informative with respect to the image. To improve this, we devise a simple auxiliary objective that incentivizes Q-BOT to ask diverse questions, thus reducing repetitions and in turn enabling A-BOT to explore a larger state space during RL i.e. be exposed to more visual concepts to talk about, and varied questions to answer. We evaluate our approach via a host of automatic metrics and human studies, and demonstrate that it leads to better dialog, i.e. dialog that is more diverse (i.e. less repetitive), consistent (i.e. has fewer conflicting exchanges), fluent (i.e. more humanlike), and detailed, while still being comparably image-relevant as prior work and ablations.

Learning to Speak and Act in a Fantasy Text Adventure Game (EMNLP, Nov 2019; Link)

Abstract:
We introduce a large-scale crowdsourced text adventure game as a research platform for studying grounded dialogue. In it, agents can perceive, emote, and act whilst conducting dialogue with other agents. Models and humans can both act as characters within the game. We describe the results of training state-of-the-art generative and retrieval models in this setting. We show that in addition to using past dialogue, these models are able to effectively use the state of the underlying world to condition their predictions. In particular, we show that grounding on the details of the local environment, including location descriptions, and the objects (and their affordances) and characters (and their previous actions) present within it allows better predictions of agent behavior and dialogue. We analyze the ingredients necessary for successful grounding in this setting, and how each of these factors relate to agents that can talk and act successfully.

Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded (ICCV, Oct 2019; Link)

Abstract:
Many vision and language models suffer from poor visual grounding – often falling back on easy-to-learn language priors rather than basing their decisions on visual concepts in the image. In this work, we propose a generic approach called Human Importance-aware Network Tuning (HINT) that effectively leverages human demonstrations to improve visual grounding. HINT encourages deep networks to be sensitive to the same input regions as humans. Our approach optimizes the alignment between human attention maps and gradient-based network importances – ensuring that models learn not just to look at but rather rely on visual concepts that humans found relevant for a task when making predictions. We apply HINT to Visual Question Answering and Image Captioning tasks, outperforming top approaches on splits that penalize over-reliance on language priors (VQA-CP and robust captioning) using human attention demonstrations for just 6% of the training data.

Task-Driven Modular Networks for Zero-Shot Compositional Learning (ICCV, Oct 2019; Link)

Abstract:
One of the hallmarks of human intelligence is the ability to compose learned knowledge into novel concepts which can be recognized without a single training example. In contrast, current state-of-the-art methods require hundreds of training examples for each possible category to build reliable and accurate classifiers. To alleviate this striking difference in efficiency, we propose a task-driven modular architecture for compositional reasoning and sample efficient learning. Our architecture consists of a set of neural network modules, which are small fully connected layers operating in semantic concept space. These modules are configured through a gating function conditioned on the task to produce features representing the compatibility between the input image and the concept under consideration. This enables us to express tasks as a combination of subtasks and to generalize to unseen categories by reweighting a set of small modules. Furthermore, the network can be trained efficiently as it is fully differentiable and its modules operate on small sub-spaces. We focus our study on the problem of compositional zero-shot classification of object-attribute categories. We show in our experiments that current evaluation metrics are flawed as they only consider unseen object-attribute pairs. When extending the evaluation to the generalized setting which accounts also for pairs seen during training, we discover that naïve baseline methods perform similarly or better than current approaches. However, our modular network is able to outperform all existing approaches on two widely-used benchmark datasets.

Habitat: A Platform for Embodied AI Research (ICCV, Oct 2019; Link)

Abstract:
We present Habitat, a platform for research in embodied artificial intelligence (AI). Habitat enables training embodied agents (virtual robots) in highly efficient photorealistic 3D simulation. Specifically, Habitat consists of: (i) Habitat-Sim: a flexible, high-performance 3D simulator with configurable agents, sensors, and generic 3D dataset handling. Habitat-Sim is fast – when rendering a scene from Matterport3D, it achieves several thousand frames per second (fps) running single-threaded, and can reach over 10,000 fps multi-process on a single GPU. (ii) Habitat-API: a modular high-level library for end-to- end development of embodied AI algorithms – defining tasks (e.g. navigation, instruction following, question answering), configuring, training, and benchmarking embodied agents. These large-scale engineering contributions enable us to answer scientific questions requiring experiments that were till now impracticable or 'merely' impractical. Specifically, in the context of point-goal navigation: (1) we revisit the comparison between learning and SLAM approaches from two recent works and find evidence for the opposite conclusion – that learning outperforms SLAM if scaled to an order of magnitude more experience than previous investigations, and (2) we conduct the first cross-dataset generalization experiments {train, test} × {Matterport3D, Gibson} for multiple sensors {blind, RGB, RGBD, D} and find that only agents

with depth (D) sensors generalize across datasets. We hope that our open-source platform and these findings will advance research in embodied AI.

Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives (ACL, Aug 2019; Link)

Abstract:
This paper tackles the problem of reading comprehension over long narratives where documents easily span over thousands of tokens. We propose a curriculum learning (CL) based Pointer-Generator framework for reading/sampling over large documents, enabling diverse training of the neural model based on the notion of alternating contextual difficulty. This can be interpreted as a form of domain randomization and/or generative pretraining during training. To this end, the usage of the Pointer-Generator softens the requirement of having the answer within the context, enabling us to construct diverse training samples for learning. Additionally, we propose a new Introspective Alignment Layer (IAL), which reasons over decomposed alignments using block-based self-attention. We evaluate our proposed method on the NarrativeQA reading comprehension benchmark, achieving state-of-the-art performance, improving existing baselines by 51% relative improvement on BLEU-4 and 17% relative improvement on Rouge-L. Extensive ablations confirm the effectiveness of our proposed IAL and CL components.

Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification (Workshop, Aug 2019; Link)

Abstract:
Interactions among users on social network platforms are usually positive, constructive and insightful. However, sometimes people also get exposed to objectionable content such as hate speech, bullying, and verbal abuse etc. Most social platforms have explicit policy against hate speech because it creates an environment of intimidation and exclusion, and in some cases may promote real-world violence. As users' interactions on today's social networks involve multiple modalities, such as texts, images and videos, in this paper we explore the challenge of automatically identifying hate speech with deep multimodal technologies, extending previous research which mostly focuses on the text signal alone. We present a number of fusion approaches to integrate text and photo signals. We show that augmenting text with image embedding information immediately leads to a boost in performance, while applying additional attention fusion methods brings further improvement.

How to Get Past Sesame Street: Sentence-Level Pretraining Beyond Language Modeling (ACL, Jul 2019; Link)

Abstract:
Natural language understanding has recently seen a surge of progress with the use of sentence encoders like ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2019) which are pretrained on variants of language modeling. We conduct the first large-scale systematic study of candidate pretraining tasks, comparing 19 different tasks both as alternatives and complements to language modeling. Our primary results support the use language modeling, especially when combined with pretraining on additional labeled-data tasks. However, our results are mixed across pretraining tasks and show some concerning trends: In ELMo's pretrain-then-freeze paradigm, random baselines are worryingly strong and results vary strikingly across target tasks. In addition, fine-tuning BERT on an intermediate task often negatively impacts downstream transfer. We also see modest gains from multitask training, suggesting the development of more sophisticated multitask and transfer learning techniques as an avenue for further research.

Learning from Dialogue after Deployment: Feed Yourself, Chatbot! (ACL, Jul 2019; Link)

Abstract:
The majority of conversations a dialogue agent sees over its lifetime occur after it has already been trained and deployed, leaving a vast store of potential training signal untapped. In this work, we propose the self-feeding chatbot, a dialogue agent with the ability to extract new training examples from the conversations it participates in. As our agent engages in conversation, it also estimates user satisfaction in its responses. When the conversation appears to be going well, the user's responses become new training examples to imitate. When the agent believes it has made a mistake, it asks for feedback; learning to predict the feedback that will be given improves the chatbot's dialogue abilities further. On the PersonaChat chit-chat dataset with over 131k training examples, we find that learning from dialogue with a self-feeding chatbot significantly improves performance, regardless of the amount of traditional supervision.

Unsupervised Question Answering by Cloze Translation (ACL, Jul 2019; Link)

Abstract:
Obtaining training data for Question Answering (QA) is time-consuming and resource-intensive, and existing QA datasets are only available for limited domains and languages. In this work, we explore to what extent high quality training data is actually required for Extractive QA, and investigate the possibility of unsupervised Extractive QA. We

approach this problem by first learning to generate context, question and answer triples in an unsupervised manner, which we then use to synthesize Extractive QA training data automatically. To generate such triples, we first sample random context paragraphs from a large corpus of documents and then random noun phrases or named entity mentions from these paragraphs as answers. Next we convert answers in context to "fill-in-the-blank" cloze questions and finally translate them into natural questions. We propose and compare various unsupervised ways to perform cloze-to-natural question translation, including training an unsupervised NMT model using non-aligned corpora of natural questions and cloze questions as well as a rule-based approach. We find that modern QA models can learn to answer human questions surprisingly well using only synthetic training data. We demonstrate that, without using the SQuAD training data at all, our approach achieves 56.4 F1 on SQuAD v1 (64.5 F1 when the answer is a Named entity mention), outperforming early supervised models.

The Referential Reader: A Recurrent Entity Network for Anaphora Resolution (ACL, Jul 2019; Link)

Abstract:
We present a new architecture for storing and accessing entity mentions during online text processing. While reading the text, entity references are identified, and may be stored by either updating or overwriting a cell in a fixed-length memory. The update operation implies coreference with the other mentions that are stored in the same cell; the overwrite operation causes these mentions to be forgotten. By encoding the memory operations as differentiable gates, it is possible to train the model end-to-end, using both a supervised anaphora resolution objective as well as a supplementary language modeling objective. Evaluation on a dataset of pronoun-name anaphora demonstrates strong performance with purely incremental text processing.

CraftAssist: A Framework for Dialogue-enabled Interactive Agents (Jul 2019; Link)

Abstract:
This paper describes an implementation of a bot assistant in Minecraft, and the tools and platform allowing players to interact with the bot and to record those interactions. The purpose of building such an assistant is to facilitate the study of agents that can complete tasks specified by dialogue, and eventually, to learn from dialogue interactions.

CNNs found to jump around more skillfully than RNNs:Compositional generalization in seq2seq convolutional networks (ACL, Jun 2019; Link)

Abstract:
Lake and Baroni (2018) introduced the SCAN dataset probing the ability of seq2seq models to capture compositional generalizations, such as inferring the meaning of “jump around” 0-shot from the component words. Recurrent networks (RNNs) were found to completely fail the most challenging generalization cases. We test here a convolutional network (CNN) on these tasks, reporting hugely improved performance with respect to RNNs. Despite the big improvement, the CNN has however not induced systematic rules, suggesting that the difference between compositional and non-compositional behaviour is not clear-cut.

Neural Models of Text Normalization for Speech Applications (Computational Linguistics, Jun 2019; Link)

Abstract:
Machine learning, including neural network techniques, have been applied to virtually every domain in natural language processing. One problem that has been somewhat resistant to effective machine learning solutions is text normalization for speech applications such as text-to-speech synthesis (TTS). In this application, one must decide, for example, that 123 is verbalized as one hundred twenty three in 123 pages but as one twenty three in 123 King Ave. For this task, state-of-the-art industrial systems depend heavily on hand-written language-specific grammars.

We propose neural network models which treat text normalization for TTS as a sequence-to-sequence problem, in which the input is a text token in context, and the output is the verbalization of that token. We find that the most effective model, in accuracy and efficiency, is one where the sentential context is computed once and the results of that computation are combined with the computation of each token in sequence to compute the verbalization. This model allows for a great deal of flexibility in terms of representing the context, and also allows us to integrate tagging and segmentation into the process.

These models perform very well overall, but occasionally they will predict wildly inappropriate verbalizations, such as reading 3cm as three kilometers. While rare, such verbalizations are a major issue for TTS applications. We thus use finite-state covering grammars to guide the neural models, either during training and decoding, or just during decoding, away from such “unrecoverable” errors. Such grammars can largely be learned from data.

What makes a good conversation? How controllable attributes affect human judgments (NAACL, May 2019; Link)

Abstract:
A good conversation requires balance – between simplicity and detail; staying on topic and changing it; asking questions and answering them. Although dialogue agents are commonly evaluated via human judgments of overall quality, the relationship between quality and these individual factors is less well-studied. In this work, we examine two controllable neural text generation methods, conditional training and weighted decoding, in order to control four important attributes for chitchat dialogue: repetition, specificity, response-relatedness and question-asking. We conduct a large-scale human evaluation to measure the effect of these control parameters on multi-turn interactive conversations on the PersonaChat task. We provide a detailed analysis of their relationship to high-level aspects of conversation, and show that by controlling combinations of these variables our models obtain clear improvements in human quality judgments.

Towards VQA Models That Can Read (CVPR, May 2019; Link)

Abstract:
Studies have shown that a dominant class of questions asked by visually impaired users on images of their surroundings involves reading text in the image. But today's VQA models can not read! Our paper takes a first step towards addressing this problem. First, we introduce a new "TextVQA" dataset to facilitate progress on this important problem. Existing datasets either have a small proportion of questions about text (e.g., the VQA dataset) or are too small (e.g., the VizWiz dataset). TextVQA contains 45,336 questions on 28,408 images that require reasoning about text to answer. Second, we introduce a novel model architecture that reads text in the image, reasons about it in the context of the image and the question, and predicts an answer which might be a deduction based on the text and the image or composed of the strings found in the image. Consequently, we call our approach Look, Read, Reason & Answer (LoRRA). We show that LoRRA outperforms existing state-of-the-art VQA models on our TextVQA dataset. We find that the gap between human performance and machine performance is significantly larger on TextVQA than on VQA 2.0, suggesting that TextVQA is well-suited to benchmark progress along directions complementary to VQA 2.0.

Generative Question Answering: Learning to Answer the Whole Question (ICLR, May 2019; Link)

Abstract:
Discriminative question answering models can overfit to superficial biases in datasets, because their loss function saturates when any clue makes the answer likely. We introduce

generative models of the joint distribution of questions and answers, which are trained to explain the whole question, not just to answer it. Our question answering (QA) model is implemented by learning a prior over answers, and a conditional language model to generate the question given the answer— allowing scalable and interpretable many-hop reasoning as the question is generated word-by-word. Our model achieves competitive performance with comparable discriminative models on the SQUAD and CLEVR benchmarks, indicating that it is a more general architecture for language understanding and reasoning than previous work. The model greatly improves generalisation both from biased training data and to adversarial testing data, achieving state-of-the-art results on ADVERSARIALSQUAD.

Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia (ACM CHI, May 2019; Link)

Abstract:
People with dyslexia face challenges expressing themselves in writing on social networking sites (SNSs). Such challenges come from not only the technicality of writing, but also the self-representation aspect of sharing and communicating publicly on social networking sites such as Facebook. To empower people with dyslexia-style writing to express themselves more confidently on SNSs, we designed and implemented Additional Writing Help (AWH) – a writing assistance tool to proofread text produced by users with dyslexia before they post on Facebook. AWH was powered by a neural machine translation (NMT) model that translates dyslexia style to non-dyslexia style writing. We evaluated the performance and the design of AWH through a week-long field study with 19 people with dyslexia and received highly positive feedback. Our field study demonstrated the value of providing better and more extensive writing support on SNSs, and the potential of AI for building a more inclusive Internet.

code2seq: Generating Sequences from Structured Representations of Code (ICLR, May 2019; Link)

Abstract:
The ability to generate natural language sequences from source code snippets has a variety of applications such as code summarization, documentation, and retrieval. Sequence-to-sequence (seq2seq) models, adopted from neural machine translation (NMT), have achieved state-of-the-art performance on these tasks by treating source code as a sequence of tokens. We present CODE2SEQ: an alternative approach that leverages the syntactic structure of programming languages to better encode source code. Our model represents a code snippet as the set of compositional paths in its abstract syntax tree (AST) and uses attention to select the relevant paths while decoding. We demonstrate the effectiveness of our approach for two tasks, two programming languages, and four datasets of up to 16M examples. Our model significantly outperforms previous models that were

specifically designed for programming languages, as well as state-of-the-art NMT models. An online demo of our model is available at http://code2seq.org. Our code, data and trained models are available here.

PyTorch-BigGraph: A Large-scale Graph Embedding System (SysML, Apr 2019; Link)

Abstract:
Graph embedding methods produce unsupervised node features from graphs that can then be used for a variety of machine learning tasks. Modern graphs, particularly in industrial applications, contain billions of nodes and trillions of edges, which exceeds the capability of existing embedding systems. We present PyTorch-BigGraph (PBG), an embedding system that incorporates several modifications to traditional multi-relation embedding systems that allow it to scale to graphs with billions of nodes and trillions of edges. PBG uses graph partitioning to train arbitrarily large embeddings on either a single machine or in a distributed environment. We demonstrate comparable performance with existing embedding systems on common benchmarks, while allowing for scaling to arbitrarily large graphs and parallelization on multiple machines. We train and evaluate embeddings on several large social network graphs as well as the full Freebase dataset, which contains over 100 million nodes and 2 billion edges.

No Training Required: Exploring Random Encoders for Sentence Classification (ICLR, Mar 2019; Link)

Abstract:
We explore various methods for computing sentence representations from pretrained word embeddings without any training, i.e., using nothing but random parameterizations. Our aim is to put sentence embeddings on more solid footing by 1) looking at how much modern sentence embeddings gain over random methods -- as it turns out, surprisingly little; and by 2) providing the field with more appropriate baselines going forward -- which are, as it turns out, quite strong. We also make important observations about proper experimental protocol for sentence classification evaluation, together with recommendations for future research.

PyText: A seamless path from NLP research to production (Dec 2018; Link)

Abstract:
We introduce PyText1 – a deep learning based NLP modeling framework built on PyTorch. PyText addresses the often-conflicting requirements of enabling rapid experimentation and

of serving models at scale. It achieves this by providing simple and extensible interfaces for model components, and by using PyTorch's capabilities of exporting models for inference via the optimized Caffe2 execution engine. We report our own experience of migrating experimentation and production workflows to PyText, which enabled us to iterate faster on novel modeling ideas and then seamlessly ship them at industrial scale.

Explore-Exploit: A Framework for Interactive and Online Learning

(NeurIPS, Dec 2018; Link)

Abstract:
Interactive user interfaces need to continuously evolve based on the interactions that a user has (or does not have) with the system. This may require constant exploration of various options that the system may have for the user and obtaining signals of user preferences on those. However, such an exploration, especially when the set of available options itself can change frequently, can lead to suboptimal user experiences. We present Explore-Exploit: a framework designed to collect and utilize user feedback in an interactive and online setting that minimizes regressions in end-user experience. This framework provides a suite of online learning operators for various tasks such as personalization ranking, candidate selection and active learning. We demonstrate how to integrate this framework with run-time services to leverage online and interactive machine learning out-of-the-box. We also present results demonstrating the efficiencies that can be achieved using the Explore-Exploit framework.

Do Explanations Make VQA Models More Predictable To A Human?

(EMNLP, Nov 2018; Link)

Abstract:
A rich line of research attempts to make deep neural networks more transparent by generating human-interpretable 'explanations' of their decision process, especially for interactive tasks like Visual Question Answering (VQA). In this work, we analyze if existing explanations indeed make a VQA model – its responses as well as failures – more predictable to a human. Surprisingly, we find that they do not. On the other hand, we find that human-in-the-loop approaches that treat the model as a black-box do.

Reference-less Quality Estimation of Text Simplification Systems

(INLG, Oct 2018; Link)

Abstract:
The evaluation of text simplification (TS) systems remains an open challenge. As the task has common points with machine translation (MT), TS is often evaluated using MT metrics

such as BLEU. However, such metrics require high quality reference data, which is rarely available for TS. TS has the advantage over MT of being a monolingual task, which allows for direct comparisons to be made between the simplified text and its original version. In this paper, we compare multiple approaches to reference-less quality estimation of sentence-level text simplification systems, based on the dataset used for the QATS 2016 shared task. We distinguish three different dimensions: grammaticality, meaning preservation and simplicity. We show that n-gram-based MT metrics such as BLEU and METEOR correlate the most with human judgment of grammaticality and meaning preservation, whereas simplicity is best evaluated by basic length-based metrics.

Extending Neural Generative Conversational Model using External Knowledge Sources (EMNLP, Oct 2018; Link)

Abstract:
The use of connectionist approaches in conversational agents has been progressing rapidly due to the availability of large corpora. However current generative dialogue models often lack coherence and are content poor. This work proposes an architecture to incorporate unstructured knowledge sources to enhance the next utterance prediction in chit-chat type of generative dialogue models. We focus on Sequence-to-Sequence (Seq2Seq) conversational agents trained with the Reddit News dataset, and consider incorporating external knowledge from Wikipedia summaries as well as from the NELL knowledge base. Our experiments show faster training time and improved perplexity when leveraging external knowledge.

Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition (CoRL, Oct 2018; Link)

Abstract:
In an open-world setting, it is inevitable that an intelligent agent (e.g., a robot) will encounter visual objects, attributes or relationships it does not recognize. In this work, we develop an agent empowered with visual curiosity, i.e. the ability to ask questions to an Oracle (e.g., human) about the contents in images (e.g., What is the object on the left side of the red cube?) and build visual recognition model based on the answers received (e.g., Cylinder). In order to do this, the agent must (1) understand what it recognizes and what it does not, (2) formulate a valid, unambiguous and informative language query (a question) to ask the Oracle, (3) derive the parameters of visual classifiers from the Oracle response and (4) leverage the updated visual classifiers to ask more clarified questions. Specifically, we propose a novel framework and formulate the learning of visual curiosity as a reinforcement learning problem. In this framework, all components of our agent, visual recognition module (to see), question generation policy (to ask), answer digestion module (to understand) and graph memory module (to memorize), are learned entirely end-to-end to maximize the reward derived from the scene graph obtained by the agent as a consequence of the dialog with the Oracle. Importantly, the question generation policy is disentangled from the visual

recognition system and specifics of the environment. Consequently, we demonstrate a sort of double generalization. Our question generation policy generalizes to new environments and a new pair of eyes, i.e., new visual system. Trained on a synthetic dataset, our results show that our agent learns new visual concepts significantly faster than several heuristic baselines, even when tested on synthetic environments with novel objects, as well as in a realistic environment.

TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering (KDD, Aug 2018; Link)

Abstract:
Taxonomy construction is not only a fundamental task for semantic analysis of text corpora, but also an important step for applications such as information filtering, recommendation, and Web search. Existing pattern-based methods extract hypernym-hyponym term pairs and then organize these pairs into a taxonomy. However, by considering each term as an independent concept node, they over-look the topical proximity and the semantic correlations among terms. In this paper, we propose a method for constructing topic taxonomies, wherein every node represents a conceptual topic and is defined as a cluster of semantically coherent concept terms. Our method, TaxoGen, uses term embeddings and hierarchical cluster-ing to construct a topic taxonomy in a recursive fashion. To ensure the quality of the recursive process, it consists of: (1) an adaptive spherical clustering module for allocating terms to proper levels when splitting a coarse topic into fine-grained ones; (2) a local embedding module for learning term embeddings that maintain strong discriminative power at different levels of the taxonomy. Our experiments on two real datasets demonstrate the effectiveness of TaxoGen compared with baseline methods.

Controllable Abstractive Summarization (ACL, Jul 2018; Link)

Abstract:
Current models for document summarization disregard user preferences such as the desired length, style, the entities that the user might be interested in, or how much of the document the user has already read. We present a neural summarization model with a simple but effective mechanism to enable users to specify these high level attributes in order to control the shape of the final summaries to better suit their needs. With user input, our system can produce high quality summaries that follow user preferences. Without user input, we set the control variables automatically – on the full text CNN-Dailymail dataset, we outperform state of the art abstractive systems (both in terms of F1-ROUGE1 40.38 vs. 39.53 F1-ROUGE and human evaluation).

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties (ACL, Jul 2018; Link)

Abstract:
Although much effort has recently been devoted to training high-quality sentence embeddings, we still have a poor understanding of what they are capturing. “Downstream” tasks, often based on sentence classification, are commonly used to evaluate the quality of sentence representations. The complexity of the tasks makes it however difficult to infer what kind of information is present in the representations. We introduce here 10 probing tasks designed to capture simple linguistic features of sentences, and we use them to study embeddings generated by three different encoders trained in eight distinct ways, uncovering intriguing properties of both encoders and training methods.

Hierarchical Neural Story Generation (ACL, Jul 2018; Link)

Abstract:
We explore story generation: creative systems that can build coherent and fluent passages of text about a topic. We collect a large dataset of 300K human-written stories paired with writing prompts from an online forum. Our dataset enables hierarchical story generation, where the model first generates a premise, and then transforms it into a passage of text. We gain further improvements with a novel form of model fusion that improves the relevance of the story to the prompt, and adding a new gated multi-scale self-attention mechanism to model long-range context. Experiments show large improvements over strong baselines on both automated and human evaluations. Human judges prefer stories generated by our approach to those from a strong non-hierarchical model by a factor of two to one.

Personalizing Dialogue Agents: I have a dog, do you have pets too? (ACL, Jul 2018; Link)

Abstract:
Chit-chat models are known to have several problems: they lack specificity, do not display a consistent personality and are often not very captivating. In this work we present the task of making chit-chat more engaging by conditioning on profile information. We collect data and train models to (i) condition on their given profile information; and (ii) information about the person they are talking to, resulting in improved dialogues, as measured by next utterance prediction. Since (ii) is initially unknown, our model is trained to engage its partner with personal topics, and we show the resulting dialogue can be used to predict profile information about the interlocutors.

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence (CVPR, Jun 2018; Link)

Abstract:
Deep models that are both effective and explainable are desirable in many settings; prior explainable models have been unimodal, offering either image-based visualization of attention weights or text-based generation of post-hoc justifications. We propose a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths. We collect two new datasets to define and evaluate this task, and propose a novel model which can provide joint textual rationale generation and attention visualization. Our datasets define visual and textual justifications of a classification decision for activity recognition tasks (ACT-X) and for visual question answering tasks (VQA-X). We quantitatively show that training with the textual explanations not only yields better textual justification models, but also better localizes the evidence that supports the decision. We also qualitatively show cases where visual explanation is more insightful than textual explanation, and vice versa, supporting our thesis that multimodal explanation models offer significant benefits over unimodal approaches.

Separating Self-Expression and Visual Content in Hashtag Supervision (CVPR, Jun 2018; Link)

Abstract:
The variety, abundance, and structured nature of hashtags make them an interesting data source for training vision models. For instance, hashtags have the potential to significantly reduce the problem of manual supervision and annotation when learning vision models for a large number of concepts. However, a key challenge when learning from hashtags is that they are inherently subjective because they are provided by users as a form of self-expression. As a consequence, hashtags may have synonyms (different hashtags referring to the same visual content) and may be polysemous (the same hashtag referring to different visual content). These challenges limit the effectiveness of approaches that simply treat hashtags as image-label pairs. This paper presents an approach that extends upon modeling simple image-label pairs with a joint model of images, hashtags, and users. We demonstrate the efficacy of such approaches in image tagging and retrieval experiments, and show how the joint model can be used to perform user-conditional retrieval and tagging.

QuickEdit: Editing Text & Translations by Crossing Words Out (NAACL, Jun 2018; Link)

Abstract:
We propose a framework for computer-assisted text editing. It applies to translation post-editing and to paraphrasing. Our proposal relies on very simple interactions: a human

editor modifies a sentence by marking tokens they would like the system to change. Our model then generates a new sentence which reformulates the initial sentence by avoiding marked words. The approach builds upon neural sequence-to-sequence modeling and introduces a neural network which takes as input a sentence along with change markers. Our model is trained on translation bitext by simulating post-edits. We demonstrate the advantage of our approach for translation post-editing through simulated post-edits. We also evaluate our model for paraphrasing through a user study.

Advances in Pre-Training Distributed Word Representations (LREC, May 2018; Link)

Abstract:
Many Natural Language Processing applications nowadays rely on pre-trained word representations estimated from large text corpora such as news collections, Wikipedia and Web Crawl. In this paper, we show how to train high-quality word vector representations by using a combination of known tricks that are however rarely used together. The main result of our work is the new set of publicly available pre-trained models that outperform the current state of the art by a large margin on a number of tasks.

Consequentialist Conditional Cooperation in Social Dilemmas with Imperfect Information (ICLR, Apr 2018; Link)

Abstract:
Social dilemmas, where mutual cooperation can lead to high payoffs but participants face incentives to cheat, are ubiquitous in multi-agent interaction. We wish to construct agents that cooperate with pure cooperators, avoid exploitation by pure defectors, and incentivize cooperation from the rest. However, often the actions taken by a partner are (partially) unobserved or the consequences of individual actions are hard to predict. We show that in a large class of games good strategies can be constructed by conditioning one's behavior solely on outcomes (ie. one's past rewards). We call this consequentialist conditional cooperation. We show how to construct such strategies using deep reinforcement learning techniques and demonstrate, both analytically and experimentally, that they are effective in social dilemmas beyond simple matrix games. We also show the limitations of relying purely on consequences and discuss the need for understanding both the consequences of and the intentions behind an action.

Efficient Large-Scale Multi-Modal Classification (AAAI, Feb 2018; Link)

Abstract:
While the incipient internet was largely text-based, the modern digital world is becoming increasingly multi-modal. Here, we examine multi-modal classification where one modality is discrete, e.g. text, and the other is continuous, e.g. visual representations transferred from a convolutional neural network. In particular, we focus on scenarios where we have to be able to classify large quantities of data quickly. We investigate various methods for performing multi-modal fusion and analyze their trade-offs in terms of classification accuracy and computational efficiency. Our findings indicate that the inclusion of continuous information improves performance over text-only on a range of multi-modal classification tasks, even with simple fusion methods. In addition, we experiment with discretizing the continuous features in order to speed up and simplify the fusion process even further. Our results show that fusion with discretized features outperforms text-only classification, at a fraction of the computational cost of full multimodal fusion, with the additional benefit of improved interpretability.

StarSpace: Embed All The Things! (AAAI, Feb 2018; Link)

Abstract:
We present StarSpace, a general-purpose neural embedding model that can solve a wide variety of problems: labeling tasks such as text classification, ranking tasks such as information retrieval/web search, collaborative filtering-based or content-based recommendation, embedding of multi-relational graphs, and learning word, sentence or document level embeddings. In each case the model works by embedding those entities comprised of discrete features and comparing them against each other – learning similarities dependent on the task. Empirical results on a number of tasks show that StarSpace is highly competitive with existing methods, whilst also being generally applicable to new cases where those methods are not.

Attentive Explanations: Justifying Decisions and Pointing to the Evidence (NIPS, Dec 2017; Link)

Abstract:
Deep models are the defacto standard in visual decision problems due to their impressive performance on a wide array of visual tasks. On the other hand, their opaqueness has led to a surge of interest in explainable systems. In this work, we emphasize the importance of model explanation in various forms such as visual pointing and textual justification. The lack of data with justification annotations is one of the bottlenecks of generating multimodal explanations. Thus, we propose two large-scale datasets with annotations that visually and

textually justify a classification decision for various activities, i.e. ACT-X, and for question answering, i.e. VQA-X. We also introduce a multimodal methodology for generating visual and textual explanations simultaneously. We quantitatively show that training with the textual explanations not only yields better textual justification models, but also models that better localize the evidence that support their decision.

One-Sided Unsupervised Domain Mapping (NIPS, Dec 2017; Link)

Abstract:
In unsupervised domain mapping, the learner is given two unmatched datasets A and B. The goal is to learn a mapping GAB that translates a sample in A to the analog sample in B. Recent approaches have shown that when learning simultaneously both GAB and the inverse mapping GBA, convincing mappings are obtained. In this work, we present a method of learning GAB without learning GBA. This is done by learning a mapping that maintains the distance between a pair of samples. Moreover, good mappings are obtained, even by maintaining the distance between different parts of the same sample before and after mapping. We present experimental results that the new method not only allows for one sided mapping learning, but also leads to preferable numerical results over the existing circularity-based constraint. Our entire code is here.

Gradient Episodic Memory for Continual Learning (NIPS, Dec 2017; Link)

Abstract:
One major obstacle towards AI is the poor ability of models to solve new problems quicker, and without forgetting previously acquired knowledge. To better understand this issue, we study the problem of continual learning, where the model observes, once and one by one, examples concerning a sequence of tasks. First, we propose a set of metrics to evaluate models learning over a continuum of data. These metrics characterize models not only by their test accuracy, but also in terms of their ability to transfer knowledge across tasks. Second, we propose a model for continual learning, called Gradient Episodic Memory (GEM) that alleviates forgetting, while allowing beneficial transfer of knowledge to previous tasks. Our experiments on variants of the MNIST and CIFAR-100 datasets demonstrate the strong performance of GEM when compared to the state-of-the-art.

Unbounded Cache Model for Online Language Modeling with Open Vocabulary (NIPS, Dec 2017; Link)

Abstract:

Recently, continuous cache models were proposed as extensions to recurrent neural network language models, to adapt their predictions to local changes in the data distribution. These models only capture the local context, of up to a few thousands tokens. In this paper, we propose an extension of continuous cache models, which can scale to larger contexts. In particular, we use a large scale non-parametric memory component that stores all the hidden activations seen in the past. We leverage recent advances in approximate nearest neighbor search and quantization algorithms to store millions of representations while searching them efficiently. We conduct extensive experiments showing that our approach significantly improves the perplexity of pre-trained language models on new distributions, and can scale efficiently to much larger contexts than previously proposed local cache models.

Evaluating Visual Conversational Agents via Cooperative Human-AI Games (HCOMP, Oct 2017; Link)

Abstract:
As AI continues to advance, human-AI teams are inevitable. However, progress in AI is routinely measured in isolation, without a human in the loop. It is crucial to benchmark progress in AI, not just in isolation, but also in terms of how it translates to helping humans perform certain tasks, i.e., the performance of human-AI teams. In this work, we design a cooperative game – GuessWhich – to measure human-AI team performance in the specific context of the AI being a visual conversational agent. GuessWhich involves live interaction between the human and the AI. The AI, which we call ALICE, is provided an image which is unseen by the human. Following a brief description of the image, the human questions ALICE about this secret image to identify it from a fixed pool of images. We measure performance of the human-ALICE team by the number of guesses it takes the human to correctly identify the secret image after a fixed number of dialog rounds with ALICE. We compare performance of the human-ALICE teams for two versions of ALICE. Our human studies suggest a counterintuitive trend – that while AI literature shows that one version outperforms the other when paired with an AI questioner bot, we find that this improvement in AI-AI performance does not translate to improved human-AI performance. This suggests a mismatch between benchmarking of AI in isolation and in the context of human-AI teams.

Learning to Reason: End-to-End Module Networks for Visual Question Answering (ICCV, Oct 2017; Link)

Abstract:
Natural language questions are inherently compositional, and many are most easily answered by reasoning about their decomposition into modular sub-problems. For example, to answer “is there an equal number of balls and boxes?” we can look for balls, look for boxes, count them, and compare the results. The recently proposed Neural Module Network (NMN) architecture implements this approach to question answering by parsing questions into linguistic substructures and assembling question-specific deep networks from smaller modules that each solve one subtask. However, existing NMN implementations rely on brittle

off-the-shelf parsers, and are restricted to the module configurations proposed by these parsers rather than learning them from data. In this paper, we propose End-to-End Module Networks (N2NMNs), which learn to reason by directly predicting instance-specific network layouts without the aid of a parser. Our model learns to generate network structures (by imitating expert demonstrations) while simultaneously learning network parameters (using the downstream task loss). Experimental results on the new CLEVR dataset targeted at compositional question answering show that N2NMNs achieve an error reduction of nearly 50% relative to state-of-the-art attentional approaches, while discovering interpretable network architectures specialized for each question.

Inferring and Executing Programs for Visual Reasoning (ICCV, Oct 2017; Link)

Abstract:
Existing methods for visual reasoning attempt to directly map inputs to outputs using black-box architectures without explicitly modeling the underlying reasoning processes. As a result, these black-box models often learn to exploit biases in the data rather than learning to perform visual reasoning. Inspired by module networks, this paper proposes a model for visual reasoning that consists of a program generator that constructs an explicit representation of the reasoning process to be performed, and an execution engine that executes the resulting program to produce an answer. Both the program generator and the execution engine are implemented by neural networks, and are trained using a combination of backpropagation and REINFORCE. Using the CLEVR benchmark for visual reasoning, we show that our model significantly outperforms strong baselines and generalizes better in a variety of settings.

Learning Visual N-Grams from Web Data (ICCV, Oct 2017; Link)

Abstract:
Real-world image recognition systems need to recognize tens of thousands of classes that constitute a plethora of visual concepts. The traditional approach of annotating thousands of images per class for training is infeasible in such a scenario, prompting the use of webly supervised data. This paper explores the training of image-recognition systems on large numbers of images and associated user comments, without using manually labeled images. In particular, we develop visual n-gram models that can predict arbitrary phrases that are relevant to the content of an image. Our visual n-gram models are feed-forward convolutional networks trained using new loss functions that are inspired by n-gram models commonly used in language modeling. We demonstrate the merits of our models in phrase prediction, phrase-based image retrieval, relating images and captions, and zero-shot transfer.

Deal or No Deal? End-to-End Learning for Negotiation Dialogues (EMNLP, Sep 2017; Link)

Abstract:
Much of human dialogue occurs in semi-cooperative settings, where agents with different goals attempt to agree on common decisions. Negotiations require complex communication and reasoning skills, but success is easy to measure, making this an interesting task for AI. We gather a large dataset of human-human negotiations on a multi-issue bargaining task, where agents who cannot observe each other's reward functions must reach an agreement (or a deal) via natural language dialogue. For the first time, we show it is possible to train end-to-end models for negotiation, which must learn both linguistic and reasoning skills with no annotated dialogue states. We also introduce dialogue rollouts, in which the model plans ahead by simulating possible complete continuations of the conversation, and find that this technique dramatically improves performance. Our code and dataset are publicly available.

Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection (EMNLP, Sep 2017; Link)

Abstract:
The ubiquity of metaphor in our everyday communication makes it an important problem for natural language understanding. Yet, the majority of metaphor processing systems to date rely on hand-engineered features and there is still no consensus in the field as to which features are optimal for this task. In this paper, we present the first deep learning architecture designed to capture metaphorical composition. Our results demonstrate that it outperforms the existing approaches in the metaphor identification task.

Enriching Word Vectors with Subword Information (ACL, Jul 2017; Link)

Abstract:
Continuous word representations, trained on large unlabeled corpora are useful for many natural language processing tasks. Popular models that learn such representations ignore the morphology of words, by assigning a distinct vector to each word. This is a limitation, especially for languages with large vocabularies and many rare words. In this paper, we propose a new approach based on the skipgram model, where each word is represented as a bag of character n-grams.

Reading Wikipedia to Answer Open-Domain Questions (ACL, Jul 2017; Link)

Abstract:
This paper proposes to tackle open- domain question answering using Wikipedia as the unique knowledge source: the answer to any factoid question is a text span in a Wikipedia article. This task of machine reading at scale combines the challenges of document retrieval (finding the relevant articles) with that of machine comprehension of text (identifying the answer spans from those articles). Our approach combines a search component based on bigram hashing and TF-IDF matching with a multi-layer recurrent neural network model trained to detect answers in Wikipedia paragraphs. Our experiments on multiple existing QA datasets indicate that (1) both modules are highly competitive with respect to existing counterparts and (2) multitask learning using distant supervision on their combination is an effective complete system on this challenging task.

Automatic Rule Extraction from Long Short Term Memory Networks (ICLR, Apr 2017; Link)

Abstract:
Although deep learning models have proven effective at solving problems in natural language processing, the mechanism by which they come to their conclusions is often unclear. As a result, these models are generally treated as black boxes, yielding no insight of the underlying learned patterns. In this paper we consider Long Short Term Memory networks (LSTMs) and demonstrate a new approach for tracking the importance of a given input to the LSTM for a given output. By identifying consistently important patterns of words, we are able to distill state of the art LSTMs on sentiment analysis and question answering into a set of representative phrases. This representation is then quantitatively validated by using the extracted phrases to construct a simple, rule-based classifier which approximates the output of the LSTM.

Improving Neural Language Models with a Continuous Cache (ICLR, Apr 2017; Link)

Abstract:
We propose an extension to neural network language models to adapt their prediction to the recent history. Our model is a simplified version of memory augmented networks, which stores past hidden activations as memory and accesses them through a dot product with the current hidden activation. This mechanism is very efficient and scales to very large memory sizes. We also draw a link between the use of external memory in neural network and cache models used with count based language models. We demonstrate on several language

model datasets that our approach performs significantly better than recent memory augmented networks.

Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service (CSCW, Feb 2017; Link)

Abstract:
We designed and deployed automatic alt-text (AAT), a system that applies computer vision technology to identify faces, objects, and themes from photos to generate photo alt-text for screen reader users on Facebook. We designed our system through iterations of prototyping and in-lab user studies. Our lab test participants had a positive reaction to our system and an enhanced experience with Facebook photos. We also evaluated our system through a two-week field study as part of the Facebook iOS app for 9K VoiceOver users. We randomly assigned them into control and test groups and collected two weeks of activity data and their survey feedback. The test group reported that photos on Facebook were easier to interpret and more engaging, and found Facebook more useful in general. Our system demonstrates that artificial intelligence can be used to enhance the experience for visually impaired users on social networking sites (SNSs), while also revealing the challenges with designing automated assistive technology in a SNS context.

Blogs

Meta and Microsoft Introduce the Next Generation of Llama (July 2023; Link)

- Today, we're introducing the availability of Llama 2, the next generation of our open source large language model.
- Llama 2 is free for research and commercial use.
- This release includes model weights and starting code for pretrained and fine-tuned Llama language models (Llama Chat, Code Llama) — ranging from 7B to 70B parameters.
- Microsoft and Meta are expanding their longstanding partnership, with Microsoft as the preferred partner for Llama 2.
- We're opening access to Llama 2 with the support of a broad set of companies and people across tech, academia, and policy who also believe in an open innovation approach to today's AI technologies.
- We're committed to building responsibly and are providing resources to help those who use Llama 2 do so too.

Introducing LLaMA: A foundational, 65-billion-parameter large language model (Feb 2023; Link)

As part of Meta's commitment to open science, today we are publicly releasing LLaMA (Large Language Model Meta AI), a state-of-the-art foundational large language model designed to help researchers advance their work in this subfield of AI. Smaller, more performant models such as LLaMA enable others in the research community who don't have access to large amounts of infrastructure to study these models, further democratizing access in this important, fast-changing field.

Training smaller foundation models like LLaMA is desirable in the large language model space because it requires far less computing power and resources to test new approaches, validate others' work, and explore new use cases. Foundation models train on a large set of unlabeled data, which makes them ideal for fine-tuning for a variety of tasks. We are making LLaMA available at several sizes (7B, 13B, 33B, and 65B parameters) and also sharing a LLaMA model card that details how we built the model in keeping with our approach to Responsible AI practices.

Over the last year, large language models — natural language processing (NLP) systems with billions of parameters — have shown new capabilities to generate creative text, solve mathematical theorems, predict protein structures, answer reading comprehension questions, and more. They are one of the clearest cases of the substantial potential benefits AI can offer at scale to billions of people.

Atlas: Few-shot learning with retrieval augmented language models (Jan 2023; Link)

Atlas is a retrieval-augmented language model pretrained on unlabeled data that exhibits few-shot abilities on knowledge-intensive tasks, such as Q&A and fact-checking. This new model builds on several recent research projects from FAIR (Fundamental AI Research). For retrieval, Atlas uses a dense retriever, based on a bi-encoder architecture, and is initialized with the contriever model [2]. The reader is based on a sequence-to-sequence model, which is initialized with the T5 model and uses the FiD [3] architecture to efficiently process a large number of retrieved documents. These two components — the retriever and the reader — are then jointly pre-trained using a MLM (masked language modeling) task. Knowledge distillation [4,5] is used to train the retriever component, using signals from the reader.

We released the code for our Atlas project [1] on GitHub, as well as pretrained Atlas model checkpoints, an index, and Wikipedia corpora. We present how to build a Q&A system that is trained using 100 examples and that has a small memory footprint thanks to our codebase's usability features.

How Meta uses AI to better understand people's ages on our platforms (Jan 2023; Link)

Providing age-appropriate experiences for the billions of people who use our services around the world is an important element of what we do. Understanding how old someone is underpins these efforts, but it's not an easy task. Finding new and better ways to understand people's ages online is an industry wide challenge. For large-scale companies like Meta, artificial intelligence (AI) is one of the best tools we have to help us tackle these types of challenges at scale.

For instance, we already use AI to help detect harmful misinformation, hate speech, and manipulated images. As a company, we continuously invest in AI to help keep people safe on our platforms. In particular, we invest heavily in research and technology to better understand people's ages across our platforms. Today, we're sharing new advancements for our adult classifier — an AI model we've developed to help detect whether someone is a teen or an adult.

To develop our adult classifier, we first train an AI model on signals such as profile information, like when a person's account was created and interactions with other profiles and content. For example, people in the same age group tend to interact similarly with certain types of content. From those signals, the model learns to make calculations about whether someone is an adult or a teen.

AI translates Hokkien, an unwritten language, for the first time (Oct 2022; Link)

Building an AI speech translation system for Hokkien was no easy task. These tools are usually trained on large quantities of text. But for Hokkien, there is no widely known standard writing system. Furthermore, Hokkien is what's known as an underresourced language, which means there isn't much paired speech data available in comparison with, say, Spanish or English. Also, with few human English-to-Hokkien translators, it was difficult to collect and annotate data to train the model.

To get around these problems, Meta researchers used text written in Mandarin, which is similar to Hokkien. The team also worked closely with Hokkien speakers to ensure that the translations were correct. "Our team first translated English or Hokkien speech to Mandarin text, and then translated it to Hokkien or English — both with human annotators and automatically," said Meta researcher Juan Pino. "They then added the paired sentences to the data used to train the AI model."

200 languages within a single AI model: A breakthrough in high-quality machine translation (July 2022; Link)

Meta AI has built a single AI model, NLLB-200, that is the first to translate across 200 different languages with state-of-the-art quality that has been validated through extensive evaluations for each of them.

We've also created a new evaluation dataset, FLORES-200, and measured NLLB-200's performance in each language to confirm that the translations are high quality. NLLB-200 exceeds the previous state of the art by an average of 44 percent.

We're now using modeling techniques and learnings from the project to improve and extend translations on Facebook, Instagram, and Wikipedia.

We're open-sourcing NLLB-200 models, FLORES-200, model training code, and code for re-creating the training dataset in order to help other researchers improve their translation tools and build on our work.

A Facebook-scale simulator to detect harmful behaviors (Jul 2020; Link)

For large-scale social networks, testing a proposed code update or new feature is a complex and challenging task. In person and also online, people act and interact with one another in

ways that are sometimes difficult for traditional algorithms to model or replicate. People's behavior evolves and adapts over time and is different from one geography to the next, which makes it difficult to anticipate all the ways an individual or an entire community might respond to even a small change in their environment.

To improve software testing for these complex environments — particularly in product areas related to safety, security, and privacy — Facebook researchers have developed Web-Enabled Simulation (WES). WES is a new method for building the first highly realistic, large-scale simulations of complex social networks. It has three important aspects:

- It uses machine learning to train bots to realistically simulate the behaviors of real people on a social media platform.
  Bots are trained to interact with each other using the same infrastructure as real users, so they can send messages to other bots, comment on bots' posts or publish their own, or make friend requests to other bots. Bots cannot engage with real users and their behavior cannot have any impact on real users or their experiences on the platform.

- WES is able to automate interactions between thousands or even millions of bots. We are using a combination of online and offline simulation, training bots with anything from simple rules and supervised machine learning to more sophisticated reinforcement learning. This blend gives us a spectrum of simulation characteristics that trade engineering concerns, such as speed, scale, and realism; different use cases require different engineering trade-offs along this spectrum for maximum efficiency and effectiveness.

- WES deploys these bots on the platform's actual production code base. The bots can interact with one another but are isolated from real users. This real-infrastructure simulation ensures that the bots' actions are faithful to the effects that would be witnessed by real people using the platform.

The WES approach can automatically explore complicated scenarios in a simulated environment. While the project is in a research-only stage at the moment, the hope is that one day it will help us improve our services and spot potential reliability or integrity issues before they affect real people using the platform. With WES, we are also developing the ability to answer counterfactual and what-if questions with scalability, realism, and experimental control.

Creating a dataset and a challenge for deepfakes (Sep 2019; Link)

Data sets and benchmarks have been some of the most effective tools to speed progress in AI. Our current renaissance in deep learning has been fueled in part by the ImageNet benchmark. Recent advances in natural language processing have been hastened by the GLUE and SuperGLUE benchmarks.

“Deepfake” techniques, which present realistic AI-generated videos of real people doing and saying fictional things, have significant implications for determining the legitimacy of information presented online. Yet the industry doesn't have a great dataset or benchmark for detecting them. We want to catalyze more research and development in this area and ensure that there are better open source tools to detect deepfakes. That’s why Facebook, the Partnership on AI, Microsoft, and academics from Cornell Tech, MIT, University of Oxford, UC Berkeley, University of Maryland, College Park, and University at Albany-SUNY are coming together to build the Deepfake Detection Challenge (DFDC).

The goal of the challenge is to produce technology that everyone can use to better detect when AI has been used to alter a video in order to mislead the viewer. The Deepfake Detection Challenge will include a dataset and leaderboard, as well as grants and awards, to spur the industry to create new ways of detecting and preventing media manipulated via AI from being used to mislead others. The governance of the challenge will be facilitated and overseen by the Partnership on AI’s new Steering Committee on AI and Media Integrity, which is made up of a broad cross-sector coalition of organizations including Facebook, WITNESS, Microsoft, and others in civil society and the technology, media, and academic communities.

Facebook, Carnegie Mellon build first AI that beats pros in 6-player poker (Jul 2019; Link)

- Pluribus is the first AI bot capable of beating human experts in six-player no-limit Hold’em, the most widely played poker format in the world. This is the first time an AI bot has beaten top human players in a complex game with more than two players or two teams.

- We tested Pluribus against professional poker players, including two winners of the World Series of Poker Main Event. Pluribus won decisively.

- Pluribus succeeds because it can very efficiently handle the challenges of a game with both hidden information and more than two players. It uses self-play to teach itself how to win, with no examples or guidance on strategy.

- Pluribus uses far fewer computing resources than the bots that have defeated humans in other games.

- The bot’s success will advance AI research, because many important AI challenges involve many players and hidden information.

Neural Code Search: ML-based code search using natural language queries (Jun 2019; Link)

Engineers work best when they can easily find code examples to guide them on particular coding tasks. For some questions — for example, “How to programmatically close or hide the Android soft keyboard?” — information is readily available from popular resources like Stack Overflow. But questions specific to proprietary code or APIs (or code written in less common programming languages) need a different solution, since they are not typically discussed in those forums.

To address this need, we’ve developed a code search tool that applies natural language processing (NLP) and information retrieval (IR) techniques directly to source code text. This tool, called Neural Code Search (NCS), accepts natural language queries and returns relevant code fragments retrieved directly from the code corpus. Our premise is that with the availability of large codebases, code fragments related to a developer's query are likely to be discoverable somewhere within existing large codebases. In this blog, we introduce two models that accomplish this task:

- NCS is an unsupervised model that combines NLP and IR techniques.
- UNIF is an extension of NCS that uses a supervised neural network model to improve performance when good supervision data is available for training.

Aroma: Using machine learning for code recommendation (Apr 2019; Link)

Thousands of engineers write the code to create our apps, which serve billions of people worldwide. This is no trivial task—our services have grown so diverse and complex that the codebase contains millions of lines of code that intersect with a wide variety of different systems, from messaging to image rendering. To simplify and speed the process of writing code that will make an impact on so many systems, engineers often want a way to find how someone else has handled a similar task. We created Aroma, a code-to-code search and recommendation tool that uses machine learning (ML) to make the process of gaining insights from big codebases much easier.

Prior to Aroma, none of the existing tools fully addressed this problem. Documentation tools are not always available and can be out of date, code search tools often return myriad matching results, and it is difficult to immediately find idiomatic usage patterns. With Aroma, engineers can easily find common coding patterns without the need to manually go through dozens of code snippets, saving time and energy in their day-to-day development workflow.

In addition to deploying Aroma to our internal codebase, we also created a version of Aroma on open source projects. All examples in this post are taken from a collection of 5,000 open source Android projects on GitHub.

Facebook Open Sources ELF OpenGo (May 2018; Link)

Today, Facebook AI Research (FAIR) open sourced ELF OpenGo, an AI bot that has defeated world champion professional Go players, based on our existing ELF platform for Reinforcement Learning Research. We are releasing both the trained model and the code used to create it.

Inspired by DeepMind's work, we kicked off an effort earlier this year to reproduce their recent AlphaGoZero results using FAIR's Extensible, Lightweight Framework (ELF) for reinforcement learning research. The goal was to create an open source implementation of a system that would teach itself how to play Go at the level of a professional human player or better. By releasing our code and models we hoped to inspire others to think about new applications and research directions for this technology.

ELF OpenGo has been successful playing against both other open source bots and human Go players. We played a series of games (198 wins, 2 losses) against LeelaZero (158603eb, Apr. 25, 2018), the strongest publicly available bot, using its default settings and no pondering. We also achieved a 14 win, 0 loss record against four of the top 30 world-ranked human Go players. These games were all played using a single GPU making moves every 50 seconds, Chinese rules with 7.5 komi, and unlimited time given to human players to consider their moves. We thank the LeelaZero team for their high quality work, and our hope is that open-sourcing our bot can similarly benefit community initiatives like LeelaZero. We would additionally like to thank Mr. Kim Jiseok, Mr. Shin Jinseo, Mr. Park Yeonghun, and Mr. Choi Cheolhan of the Korean Baduk association for their eager participation, challenging our bot through a series of games.

Visual reasoning and dialog: Towards natural language conversations about visual data (Nov 2017; Link)

The broad objective of visual dialog research is to teach machines to have natural language conversations with humans about visual content. This emerging field brings together aspects of computer vision, natural language processing, and dialog systems research.

In general, dialog systems can have a spectrum of capabilities. On one end of the spectrum are task-driven chat bots you can talk to for a specific goal e.g., to book a flight. On the other end are chitchat bots that you can talk to about any topic but without a clear goal in mind. Visual dialog lies somewhere in between the two extremes. It is free-form dialog but the conversation is grounded in the content of a specific image.

While visual dialog research is in the early phases, there are many potential future use cases for this technology. For example, being able to ask a series of questions could help visually-impaired people understand images that are posted online or taken of their surroundings, or allow medical personnel to better interpret medical scans. It could also have

uses in AR/VR applications where a user could chat in natural language and work with a virtual companion who is seeing what they are seeing based on a visual common ground