Curated Advanced Papers — Deep Learning | Vision–Language & Embeddings | ML/AI, Generative & Agentic AI

Prepared: 2025-11-15 17:07 UTC

Deep Learning — Core theory & applied milestones (≈20 papers)

Efficient BackProp — Y. LeCun, L. Bottou, G. B. Orr, K.-R. Müller (1998)

Paper: http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf

Classic tutorial covering optimization, nonlinearities, initialization — essential background on training deep nets.

Dropout: A Simple Way to Prevent Neural Networks from Overfitting — Nitish Srivastava et al. (2014)

Paper: https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf | Code: https://github.com/nyu-dl/dl4cv-2017-assignments/tree/master/assignment2/dropout

Introduces dropout regularization, a simple stochastic neuron dropping method that reduces overfitting and co-adaptation.

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift — S. Ioffe, C. Szegedy (2015)

Paper: https://arxiv.org/abs/1502.03167 | Code: https://github.com/tensorflow/models/tree/master/official/legacy/image_classification/resnet

BatchNorm stabilizes training, enables larger learning rates, and reduces sensitivity to initialization; widely used.

Adam: A Method for Stochastic Optimization — Diederik P. Kingma, Jimmy Ba (2015)

Paper: https://arxiv.org/abs/1412.6980 | Code: https://github.com/adam-p/adam

Popular adaptive optimizer combining momentum and adaptive learning rates; baseline for many deep learning tasks.

Understanding the difficulty of training deep feedforward neural networks (Xavier init) — Xavier Glorot, Yoshua Bengio (2010)

Paper: http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

Introduces Xavier/Glorot initialization to keep signal variance stable across layers.

Rectified Linear Units Improve Restricted Boltzmann Machines (ReLU) — A. Krizhevsky, I. Sutskever, G. Hinton (2010)

Paper: https://papers.nips.cc/paper/2010/file/1fb3ac3a8b0d0c3b2c3c4b4b0f1b8b22-Paper.pdf

Popularizes ReLU nonlinearities for faster convergence in deep nets.

ResNet: Deep Residual Learning for Image Recognition — Kaiming He et al. (2015)

Paper: https://arxiv.org/abs/1512.03385 | Code: https://github.com/KaimingHe/deep-residual-networks

Introduces residual connections allowing very deep networks to be trained (ResNet family).

Network in Network — Min Lin, Q. Chen, S. Yan (2013)

Paper: https://arxiv.org/abs/1312.4400

Micro-architectural idea (MLP conv layers) that influenced modern conv designs.

Squeeze-and-Excitation Networks — Jie Hu et al. (2017)

Paper: https://arxiv.org/abs/1709.01507 | Code: https://github.com/hujie-frank/SENet

Channel-wise attention block that boosts performance with small cost.

Attention Is All You Need — Vaswani et al. (2017)

Paper: https://arxiv.org/abs/1706.03762 | Code: https://github.com/tensorflow/tensor2tensor

Introduces Transformers — key architecture across modalities.

Vision Transformer (ViT) — Dosovitskiy et al. (2020)

Paper: https://arxiv.org/abs/2010.11929 | Code: https://github.com/google-research/vision_transformer
Shows pure Transformers can succeed on image tasks when scaled and pre-trained.

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks — Mingxing Tan & Quoc V. Le (2019)

Paper: https://arxiv.org/abs/1905.11946 | Code: https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet
Proposes compound scaling to balance depth/width/resolution for efficient accuracy.

Large Batch Training of Convolutional Networks — P. Goyal et al. (2017)

Paper: https://arxiv.org/abs/1706.02677
Techniques for training with very large batch sizes (linear scaling rule, warmup).

Layer Normalization — Jimmy Lei Ba et al. (2016)

Paper: https://arxiv.org/abs/1607.06450
Normalization method applied across features, important for RNNs and Transformers.

Weight Decay, Regularization and Generalization in Deep Nets — Various key references (1992)

Paper: https://link.springer.com/chapter/10.1007/3-540-55719-9_3
Classic theory on regularization with implications for modern deep learning.

SimCLR: A Simple Framework for Contrastive Learning of Visual Representations — Ting Chen et al. (2020)

Paper: https://arxiv.org/abs/2002.05709 | Code: https://github.com/google-research/simclr
SimCLR shows contrastive learning can produce strong representations without labels.

BYOL: Bootstrap Your Own Latent — Jean-Bastien Grill et al. (2020)

Paper: https://arxiv.org/abs/2006.07733 | Code: https://github.com/deepmind/deepmind-research/tree/master/byol
Self-supervised method that avoids negative samples; strong representation learning.

DINO: Self-Distillation with No Labels — Mathilde Caron et al. (2021)

Paper: https://arxiv.org/abs/2104.14294 | Code: https://github.com/facebookresearch/dino
Shows ViTs can learn good features without supervision using self-distillation.

Stochastic Depth and DropPath — G. Huang et al. (2016)

Paper: https://arxiv.org/abs/1603.09382
Training-time layer dropping to regularize very deep networks.

Understanding Generalization in Deep Learning — Zhang et al., Neyshabur et al., etc. (2016-2019)

Paper: https://arxiv.org/abs/1611.03530
Insights into optimization and generalization puzzles of deep nets.

Vision–Language Models & Embeddings (≈20 papers)

CLIP: Learning Transferable Visual Models From Natural Language Supervision — Radford et al. (OpenAI) (2021)

Paper: https://arxiv.org/abs/2103.00020 | Code: https://github.com/openai/CLIP
Contrastive pretraining linking images and text; strong zero-shot transfer across vision tasks.

ALIGN: Scaling Up Visual and Language Representation Learning — Jia et al. (Google) (2021)

Paper: https://arxiv.org/abs/2102.05918 | Code: https://github.com/google-research/vision_transformer
Large-scale image-text contrastive learning using noisy web alt-text; CLIP-like results at scale.

ViLBERT: Pretraining Task-Agnostic V+L Representations — Lu et al. (2019)

Paper: https://arxiv.org/abs/1908.02265 | Code: https://github.com/facebookresearch/vilbert-multi-task
Two-stream model processing image regions and text with co-attention; strong VQA and captioning.

LXMERT: Learning Cross-Modality Encoder Representations from Transformers — Tan & Bansal (2019)

Paper: https://arxiv.org/abs/1908.07490 | Code: https://github.com/airsplay/lxmert
Cross-modality pretraining for vision-and-language tasks with object features.

UNITER: UNiversal Image-TExt Representation — Chen et al. (2019)

Paper: https://arxiv.org/abs/1909.11740 | Code: https://github.com/ChenRocks/UNITER
Unified V+L pretraining combining multiple objectives; strong results on downstream tasks.

ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision — Kim et al. (2021)

Paper: https://arxiv.org/abs/2102.03334 | Code: https://github.com/dandelin/vilt
An end-to-end ViT-based approach that operates on image patches directly, simplifying pipelines.

ALBEF: Align Before Fuse — Li et al. (2021)

Paper: https://arxiv.org/abs/2107.07651 | Code: https://github.com/salesforce/ALBEF
Momentum distillation and image-text contrastive alignment before fusion improves V+L pretraining.

BLIP: Bootstrapping Language–Image Pre-training — Li et al. (2022)

Paper: https://arxiv.org/abs/2201.12086 | Code: https://github.com/salesforce/BLIP
Unified model for image captioning and VQA using bootstrapped pretraining strategies.

Flamingo: a Visual Language Model for Few-Shot Learning — DeepMind (2022)

Paper: https://arxiv.org/abs/2204.14198 | Code: https://github.com/deepmind/flamingo
Cross-modal few-shot learner with gated cross-attention between frozen image and language models.

Oscar: Object-Semantics Aligned Pre-training — Li et al. (2020)

Paper: https://arxiv.org/abs/2004.06165 | Code: https://github.com/microsoft/Oscar
Uses object tags as anchors to align image regions and text.

BLIP-2: Bootstrapped Language-Image Pretraining 2 — Li et al. (2023)

Paper: https://arxiv.org/abs/2301.12597 | Code: https://github.com/salesforce/BLIP
Bridges large frozen LLMs and vision encoders for strong V+L capabilities with low compute.

FLAVA: A Foundational Vision-Language Model — Singh et al. (Meta) (2021)

Paper: https://arxiv.org/abs/2112.04482 | Code: https://github.com/facebookresearch/FLAVA
Multimodal pretraining covering image-only, text-only, and image-text tasks.

VisualBERT: A Simple and Performant Baseline for V+L — Li et al. (2019)

Paper: https://arxiv.org/abs/1908.03557 | Code: https://github.com/uclanlp/visualbert
Joins image region features and BERT for simple VQA/captioning baselines.

OpenCLIP: Open Reproduction of CLIP — Various (2021)

Paper: https://github.com/mlfoundations/open_clip | Code: https://github.com/mlfoundations/open_clip
Community reproduction enabling research with CLIP-like models.

FLAMINGO / other few-shot V+L architectures — DeepMind/others (2022)

Paper: https://arxiv.org/abs/2204.14198
Few-shot multimodal learners built from frozen vision and language components.

DALL·E: Zero-Shot Text-to-Image Generation — Ramesh et al. (OpenAI) (2021)

Paper: https://arxiv.org/abs/2102.12092 | Code: https://github.com/openai/DALL-E
Transformer-based text-to-image synthesis demonstrating zero-shot capabilities.

PaLI / PaLI-3 and large multimodal models — Google Research (2022-2024)

Paper: https://arxiv.org/abs/2209.06794
Large-scale multimodal models for multilingual vision-language tasks.

ALUM / Image-Text Embedding techniques (FILIP, CLIP variants) — Various (2021-2023)

Paper: https://arxiv.org/abs/2101.00027

Fine-grained localization and alignment techniques for image-text embeddings.

Multimodal Chain-of-Thought and CoT variants — Various 2023-2024 (2023)

Paper: https://arxiv.org/search/?query=multimodal+chain+of+thought

Emerging research on chain-of-thought reasoning across modalities.

BLIP / ALBEF / CLIP family — code & checkpoints (repos) — Various (2021-2023)

Paper: https://github.com/salesforce/BLIP

Important repos to reproduce modern V+L experiments.

ML/AI, Generative AI & Agentic AI (≈20 papers)

Scaling Laws for Neural Language Models — Kaplan et al. (2020)

Paper: https://arxiv.org/abs/2001.08361

Defines empirical scaling laws relating model size, dataset size, and compute to performance.

Training Compute-Optimal Large Language Models (Chinchilla) — Hoffmann et al. (2022)

Paper: https://arxiv.org/abs/2203.15556

Shows optimal tradeoff between model size and dataset size; recommends training smaller models on more data.

Chain of Thought Prompting Elicits Reasoning in Large Language Models — Wei et al. (2022)

Paper: https://arxiv.org/abs/2201.11903

Shows that prompting LLMs to produce step-by-step reasoning dramatically improves multi-step problem solving.

Toolformer: Language Models Can Teach Themselves to Use Tools — Schick et al. (2023)

Paper: https://arxiv.org/abs/2302.04761 | Code: https://github.com/clare-ml/Toolformer

Automatic fine-tuning to call external tools (APIs) improving factuality and capabilities.

ReAct: Synergizing Reasoning and Acting in Language Models — Yao et al. (2022)

Paper: https://arxiv.org/abs/2210.03629 | Code: https://github.com/allenai/reaction-paper

Proposes combining reasoning traces (chain-of-thought) with action calls (tool use) for agents.

WebGPT: Browser-assisted question answering with human preferences — Nakano et al. (OpenAI) (2021)

Paper: https://arxiv.org/abs/2112.09332

LLM augmented with a web-browsing tool + human preferences to improve answer quality and citation.

Auto-GPT & BabyAGI — community agent projects — Various (2023)

Paper: https://github.com/Significant-Gravitas/Auto-GPT | Code: https://github.com/Significant-Gravitas/Auto-GPT

Community-driven autonomous agent frameworks chaining LLM calls and tools.

Generative Agents: Interactive Simulacra of Human Behavior — Park et al. (2023)

Paper: https://arxiv.org/abs/2304.03442 | Code: https://github.com/GenerativeAgents

Simulated believable human agents using memory, planning and LLMs; shows emergent social behavior.

Sparks of AGI? Evaluating Emergent Abilities of LLMs — Wei et al. (2022)

Paper: https://arxiv.org/abs/2206.07682

Studies sudden emergence of abilities when scaling LLMs; helps understand model behavior at scale.

Retrieval-Augmented Generation (RAG) — Lewis et al. (2020)

Paper: https://arxiv.org/abs/2005.11401 | Code: https://github.com/huggingface/transformers/tree/main/examples/research_projects/rag

Combines retrieval with generation for factual, grounded answers.

MuZero: Mastering Go, chess, shogi and Atari without rules — Schrittwieser et al. (2020)

Paper: https://arxiv.org/abs/1911.08265 | Code: https://github.com/werner-duvaud/muzero

Learns a model for planning via MCTS without game rules; landmark in model-based RL.

Stable Diffusion — Rombach et al. (2022)

Paper: https://arxiv.org/abs/2112.10752 | Code: https://github.com/CompVis/stable-diffusion

Latent diffusion enabling efficient, high-quality image generation on consumer hardware.

Imagen: Text-to-Image Diffusion Models — Saharia et al. (2022)

Paper: https://arxiv.org/abs/2205.11487

High-fidelity text-to-image diffusion model combining large text encoder and cascading diffusion.

DALL·E 2 / GLIDE family — OpenAI/others (2022)

Paper: https://arxiv.org/abs/2112.10741

Text-guided image generation using diffusion with classifier-free guidance and upsamplers.

Foundation Models Survey — Bommasani et al. (2021)

Paper: https://arxiv.org/abs/2108.07258

Comprehensive survey on foundation models and their implications.

Tool use, memory, and planning for agents (research overview) — Various (2022-2024)

Paper: https://arxiv.org/search/?query=agentic+ai

Rapidly growing research area—papers cover tool use, episodic memory, long-horizon planning.

Evaluation & Alignment: Red-teaming and Safety papers — Various (2020-2024)

Paper: https://arxiv.org/search/?query=alignment+language+models

Research on evaluating and aligning models and agentic behavior to safe objectives.

RLAIF / RLHF: Reinforcement learning from human preferences — Christiano et al. (2017)

Paper: https://arxiv.org/abs/1706.03741

Training models (agents) using human preference feedback—foundation for aligned LLMs.

Emergent Tool Use in LLMs and Multi-step programs — Various 2023 (2023)

Paper: https://arxiv.org/search/?query=tool+use+large+language+models

Studies showing LLMs can be chained or prompted to perform multi-step tool-using behaviors.

Survey: Autonomous Agents with LLMs — Various 2023-2024 ()

Paper: https://arxiv.org/search/?query=autonomous+agent+language+models

Collections of design patterns, benchmarks, and frameworks for agentic AI.