Dominion Akinrotimi
Data Scientist and Researcher

The Hidden work you do for AI

Dominion Akinrotimi
Data Scientist and Researcher

CAPTCHA isn't just security

When you solve CAPTCHA puzzles, you're not just proving you're human but you're helping improve **OCR and image recognition systems**.

the field of Artificial Intelligence. We introduce two families of AI problems that can be used to construct CAPTCHAs and we show that solutions to such problems can be used for steganographic communication. CAPTCHAs based on these AI problem families, then, imply a win-win situation: either the problems remain unsolved and there is a way to differentiate humans from computers, or the problems are solved and there is a way to communicate covertly on some channels.

Source: Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). CAPTCHA: Using Hard AI Problems for Security. International Conference on the Theory and Applications of Cryptographic Techniques.

Dominion Akinrotimi
Data Scientist and Researcher

Your scrolling is training data

When you like, pause, or rewatch content, you're not just scrolling, you're generating **behavioral data** used to train recommendation algorithms.

nation layer to match the softmax dimension of 256

- Depth 1: 256 ReLU
- Depth 2: 512 ReLU → 256 ReLU
- Depth 3: 1024 ReLU → 512 ReLU → 256 ReLU
- Depth 4: 2048 ReLU → 1024 ReLU → 512 ReLU → 256 ReLU

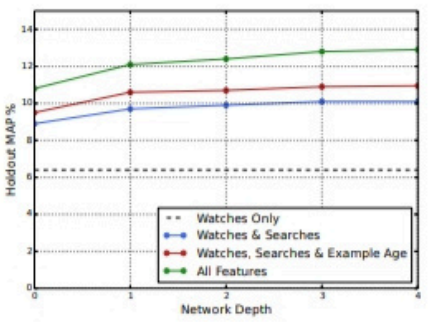


Figure 6: Features beyond video embeddings improve holdout Mean Average Precision (MAP) and layers of depth add expressiveness so that the model can effectively use these additional features by modeling their interaction.

4. RANKING

The primary role of ranking is to use impression data to specialize and calibrate candidate predictions for the particular user interface. For example, a user may watch a given

being scored rather than the millions scored in candidate generation. Ranking is also crucial for ensembling different candidate sources whose scores are not directly comparable.

We use a deep neural network with similar architecture as candidate generation to assign an independent score to each video impression using logistic regression (Figure 7). The list of videos is then sorted by this score and returned to the user. Our final ranking objective is constantly being tuned based on live A/B testing results but is generally a simple function of expected watch time per impression. Ranking by click-through rate often promotes deceptive videos that the user does not complete ("clickbait") whereas watch time better captures engagement [13, 25].

4.1 Feature Representation

Our features are segregated with the traditional taxonomy of categorical and continuous/ordinal features. The categorical features we use vary widely in their cardinality - some are binary (e.g. whether the user is logged-in) while others have millions of possible values (e.g. the user's last search query). Features are further split according to whether they contribute only a single value ("univalent") or a set of values ("multivalent"). An example of a univalent categorical feature is the video ID of the impression being scored, while a corresponding multivalent feature might be a bag of the last N video IDs the user has watched. We also classify features according to whether they describe properties of the item ("impression") or properties of the user/context ("query"). Query features are computed once per request while impression features are computed for each item scored.

Feature Engineering

We typically use hundreds of features in our ranking models, roughly split evenly between categorical and continuous. Despite the promise of deep learning to alleviate the burden of engineering features by hand, the nature of our raw data does not easily lend itself to be input directly into feedforward neural networks. We still expend considerable

Source: Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. Proceedings of the 10th ACM Conference on Recommender Systems.

Dominion Akinrotimi
Data Scientist and Researcher

You're part of how AI improves

When you like , dislike , or regenerate AI responses, you're not just reacting but you're contributing to **Reinforcement Learning from Human Feedback (RLHF)**, which improves model outputs over time.

example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of

Source: Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Dominion Akinrotimi
Data Scientist and Researcher

Your photos are datasets

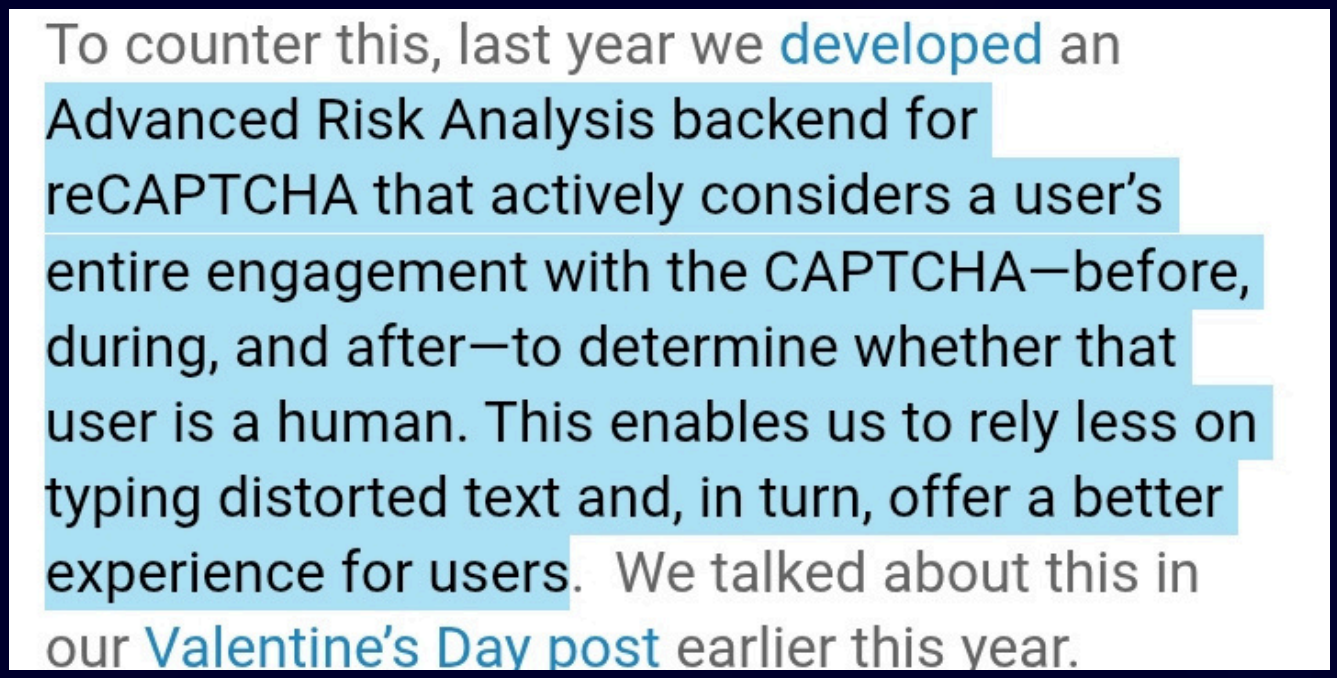
When you upload and tag photos, you're not just sharing but you're creating **labeled data** used to train computer vision systems.

Source: Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Dominion Akinrotimi
Data Scientist and Researcher

"I'm not a robot" is behavioral analysis

When you tick "I'm not a robot," you're not just verifying identity as your **mouse movements and interaction patterns** are used to distinguish humans from bots.



Source: Shet, V. (2014, December 3). No CAPTCHA reCAPTCHA. Google Security Blog. https://security.googleblog.com/2014/12/are-you-robot-introducing-no-captcha.html

Dominion Akinrotimi
Data Scientist and Researcher

You think you're using these systems.
But in small ways... **you're training them**.

Source: Synthesis of concepts from Von Ahn (2003), Covington (2016), Ouyang (2022), Deng (2009), and Shet (2014).