selbstüberwachtes Lernen

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Selbstüberwachtes Lernen konstruiert die Labels eines Trainingsdatensatzs aus den Datenpunkte selbst. Einige Merkmale eines Datenpunkts werden zurückgehalten und als dessen Label verwendet, während die übrigen als Merkmale erhalten bleiben, sodass ein gewöhnliches Problem des überwachtes Lernens entsteht, das mittels empirische Risikominimierung (ERM) auf Daten gelöst wird, die kein Annotator je berührt hat. In der natürliche Sprachverarbeitung (NLP) ist das zurückgehaltene Merkmal das nächste Token; auf diese Weise wird ein großes Sprachmodell (LLM) trainiert. Im maschinellen Sehen ist es eine Menge von Pixeln, die aus einem Bild gelöscht und anhand der sichtbar gebliebenen Bildausschnitte vorhergesagt wird.

Definition

Eine offene Bilddatenbank liefert Tausende von Fotografien für ein Stichwort wie *Katze*, und ein Web-Crawl ergibt Milliarden von Sätzen. Keines von beiden verfügt über ein Label, das angibt, was vorhergesagt werden soll. Aus jedem davon lässt sich jedoch eine Aufgabe zur Vorhersage machen: Verbirgt man das letzte Wort eines Satzes, so werden die Wörter davor zu einer Frage, deren Antwort bereits bekannt ist. Selbstüberwachtes Lernen konstruiert Labels auf diese Weise aus dem Datenpunkt selbst.

Im Allgemeinen werden einige der Merkmale eines Datenpunkts zurückgehalten und als das Label verwendet; die verbleibenden dienen weiterhin als die Merkmale. Das Ergebnis ist ein gewöhnliches Problem des überwachtes Lernens, das mittels empirische Risikominimierung (ERM) gelöst wird, und sein Trainingsdatensatz benötigt nichts weiter als die rohen Daten. Eine einzige ungelabelte Sammlung liefert so viele Trainingsdatensätze, wie es Möglichkeiten gibt, einen Teil eines Datenpunkts zu verbergen (siehe Abb. 1). Der Begriff geht auf de Sa (1993) zurück, die ihn für einen Klassifikator verwendet, dessen Label für eine Sinnesmodalität aus einer gleichzeitig auftretenden Eingabe in einer anderen abgeleitet wird.

Das zurückgehaltene Merkmal kann ein Token sein. In der natürliche Sprachverarbeitung (NLP) ist ein Datenpunkt ein Textabschnitt, seine Merkmale sind die n Tokens $v^{(1)}, \ldots, v^{(n)}$ eines Kontextfensters der Länge n, und sein Label ist das darauffolgende Token $v^{(n+1)}$. Jede Position in jedem Dokument ist ein gela-

belter Datenpunkt, sodass ein Korpus von m Tokens in der Größenordnung von m solcher Punkte liefert. Das Vorhersagen des nächsten Tokens ist die Aufgabe beim Training eines großes Sprachmodell (LLM)s wie GPT-3, eines autoregressiven Sprachmodells, d. h. eines Modells, das darauf trainiert ist, das nächste Token vorherzusagen (Brown et al., 2020, Abschn. 1). Das Maskieren eines Tokens in der Mitte und dessen Vorhersage von beiden Seiten aus hält ein anderes Merkmal desselben Datenpunkts zurück (Devlin et al., 2019).

Auf ähnliche Weise sind im maschinellen Sehen die zurückgehaltenen Merkmale Pixel: Ein Bild wird in Bildausschnitte unterteilt, die meisten davon werden gelöscht, und das Verfahren sagt die fehlenden Pixelwerte aus den verbleibenden vorher. He et al. (2022), Abschn. 1 und Abb. 2 löschen 75% der Bildausschnitte und stellen dennoch erkennbaren Inhalt wieder her.

Ein tiefes Netz, das eine der beiden Aufgaben löst, berechnet eine Verkettung $\widehat{h} = s \circ \phi$. Die Merkmalstransformation $\phi\colon \mathcal{X} \to \mathbb{R}^d$, wobei $\mathcal{X}$ die Menge der Roheingaben (Texte oder Bilder) ist, bildet einen Datenpunkt auf einen Vektor von Aktivierungen ab, und die Abbildung s sagt das zurückgehaltene Merkmal aus diesem Vektor vorher. Dieses Merkmal kann nur aus dem vorhergesagt werden, was ϕ bewahrt hat: Ein nächstes Token lässt sich nicht aus einem Vektor rekonstruieren, der den Kontext verworfen hat, und ebenso wenig ein gelöschter Bildausschnitt aus einem, der seine Umgebung ausgelassen hat. Das Trainieren der Verkettung zwingt ϕ daher dazu, diese Information zu bewahren, obwohl ϕ nirgendwo im Verlust auftritt. Diese Einschränkung legt ϕ nicht eindeutig fest: Für jede invertierbare Abbildung g von $\mathbb{R}^d$ auf sich selbst ist die Verkettung $(s \circ g^{-1}) \circ (g \circ \phi)$ gleich $s \circ \phi$ und hat somit denselben Verlust. He et al. (2022), Abschn. 3 benennen die beiden Hälften: Ihr Encoder ist ϕ und erhält nur die sichtbar gebliebenen Bildausschnitte, während ihr Decoder s ist und die Pixel aus dessen Ausgabe rekonstruiert. Was ϕ dort einschränkt, ist das Löschen des Großteils der Eingabe, nicht eine schmale Schicht wie in einem Autoencoder.

In beiden Anwendungsgebieten ist es ϕ, das übernommen wird. Die Abbildung s wird verworfen, und ein kleiner Ersatz für sie wird auf der Ausgabe von ϕ unter Verwendung der wenigen Labels angepasst, über die die spätere Aufgabe verfügt (siehe Transfer-Learning, Grundlagenmodell). Selbstüberwachtes Lernen ist daher eine Antwort auf denselben Mangel, der auch das semi-überwachtes Lernen (SSL) motiviert: Dort unterstützen ungelabelte Datenpunkte ein Problem mit wenigen Labels direkt; hier werden sie zunächst in ein eigenes gelabeltes Problem umgewandelt. Unüberwachtes Lernen ist die dritte Antwort und gibt das Label vollständig auf: Es findet Strukturen unter den Merkmalsvektoren, anstatt ein zurückgehaltenes Merkmal vorherzusagen.

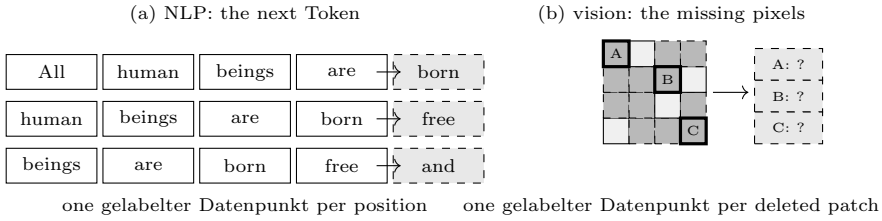


Abbildung 1: Zwei Möglichkeiten, Labels aus ungelabelten Datenpunkte zu konstruieren, wobei in beiden Fällen eine einzige Quelle viele liefert. (a) NLP: Ein Satz wird durch ein gleitendes Fenster gelesen, und jede Position ergibt einen gelabelter Datenpunkt, dessen Merkmale die Tokens im Fenster sind und dessen Label das Token danach ist (schattiert). (b) Maschinelles Sehen: Ein Bild wird in Bildausschnitte unterteilt und die meisten werden gelöscht (schattiert); jeder fehlende Bildausschnitt, von denen hier drei umrandet sind, ist ein gelabelter Datenpunkt, dessen Label seine eigenen Pixelwerte sind. Keine der beiden Aufgaben erfordert einen menschlichen Annotator

Siehe auch: Merkmal, Label, überwachtes Lernen, unüberwachtes Lernen, SSL, LLM, NLP, Token, Transfer-Learning, Grundlagenmodell.

Querverweise

Literatur

1. Brown et al. (2020). Language Models are Few-Shot Learners. In: Adv. Neural Inf. Process. Syst., pp. 1877–1901.

2. Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. 2019 Conf. North American Chapter Assoc. Computational Linguistics: Human Lang. Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

3. He et al. (2022). Masked Autoencoders Are Scalable Vision Learners. In: 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15979–15988.

4. de Sa (1993). Learning Classification with Unlabeled Data. In: Adv. Neural Inf. Process. Syst. (NIPS), pp. 112–119.